FINAL REPORT: ONR CONTRACT NO. N00014-05-C-0148

TITLE: CROSS-VALIDATION OF INDICATORS OF COGNITIVE WORKLOAD

SUBMITTED: JANUARY 10, 2006

PRINCIPAL INVESTIGATOR: SANDRA P. MARSHALL

INSTITUTION: EYETRACKING, INC.

ADDRESS: 6475 Alvarado Rd., Suite 132, San Diego, CA 92120

TELEPHONE: 619-594-2695

EMAIL: smarshall@eyetracking.com

CROSS-VALIDATION OF INDICATORS OF COGNITIVE WORKLOAD

Sandra P. Marshall
Mike Bartels

EyeTracking, Inc.
6475 Alvarado Road, Suite 132
San Diego, CA 92120

www.eyetracking.com

Final Report
Contract No. N00014-05-C-0148
December 2005

TABLE OF CONTENTS

ABSTRACT

The original 2000 AMBR project sought to evaluate how well four human performance models simulated behavior of human participants. Participants and models completed a modified version of an air traffic control task and were compared on the dimensions of performance, reaction time, and subjective workload ratings. The current study replicated the human performance findings of the previous phase of AMBR and added eye tracking analyses to enhance understanding of participants' behavior and to compare NASA TLX workload ratings with ACT-R workload predictions and ICA estimates. Examination of gaze position and patterns of eye movement provided evidence that participants adopted different visual strategies to complete the task in different display conditions and at different levels of demand. Evaluation of workload measures revealed that the three workload measures analyzed seemed to be estimating different facets of the broad concept of workload. Applicability of eye tracking analyses to understanding cognitive workload and augmenting cognitive models is discussed.

CROSS-VALIDATION OF INDICATORS OF COGNITIVE WORKLOAD

BACKGROUND

In 2000, the Air Force Research Laboratory (in association with BBN Technologies) launched the Agent-based Modeling and Behavior Representation Project (AMBR). The goal of AMBR was to evaluate and compare the accuracy of several human performance models on a single complex task. This application of multiple models to an identical task environment allowed developers to examine not only how well each model predicted behavior and performance of human participants but also which models specifically were more accurate than others and in what specific areas. The participating developers were Soar Technologies, CHI Systems, AFRL, and Carnegie Mellon University. Each provided a model to be tested against 16 human participants and compared with the other three models.

The task chosen for use in AMBR was a modified version of an air traffic control task, in which each participant acted as an air traffic controller, responsible for handling incoming and outgoing aircraft as they traversed a radar screen. The design of the task itself was particularly well suited for use in human performance modeling. In order to successfully complete all of the desired objectives of the air traffic controller, human participants and human performance models were required to complete goal-directed behaviors under time pressure. This requirement led to shifts in attention across different regions of the screen, continual prioritization of necessary actions, and management of multiple objectives simultaneously despite frequent interruptions (Deutsch and Cramer, 1998). Based on these aspects of the task, modelers were faced with the challenge of designing models that captured the strategies used by humans to process information that arrives at inconvenient and unexpected times, disrupting ongoing cognitive processes and obscuring important events and necessary actions.

Analysis dealt with model predictions of task performance, reaction time, and workload ratings. Data for each model was averaged and compared to the averages of the 16 human participants. Comparisons were made across three levels of demand of the air traffic control task and two display types: *text display*, in which task demands were conveyed through text messages alone and *color display,* in which text messages were accompanied by aircraft color-coding. Results from human participants indicated that performance suffered in the text display conditions, especially as task demand increased. Reaction times were slower in the text display condition, and this effect was more pronounced at higher levels of demand. In addition, subjective workload ratings on the NASA TLX indicated that participants felt that the text display condition at the higher levels of demand required increased cognitive effort (Tenney & Spector, 2001).

All four of the models used in comparison echoed these trends. Although some models were more accurate than others in predicting different facets of these results, the general tendency of participants to perform more poorly, more slowly, and with a higher sense of effort in the text display condition and at higher demand levels was supported by each model. Overall, the models seemed to be simulating human behavior and performance very accurately.

Despite the success of this first phase of the AMBR project, a review of the findings by an expert panel expressed some concerns about the results (Gray, 2000). First, the NASA TLX, which was used as the subjective workload measure, is somewhat suspect in its representation of actual cognitive workload. This criticism is largely based on the fact that the TLX assumes that participants are aware of and capable of interpreting their level of workload. More stressful parts of the task may not be fully reflected unless they immediately precede the ratings at the end of the simulation. In addition, some of the individual scales of the TLX are confusing, and others can not readily be applied to the AMBR task. Given the shortcomings of this subjective measure, it is clear that an objective psychophysiological measure of workload would allow researchers to assess more accurately the amount of cognitive effort expended by human participants. This objective workload measure could then be compared with estimates of workload from cognitive models such as the one provided by the Adaptive Character of Thought - Rational model (ACT-R). The current study employed a cross-validation methodology to compare the ACT-R workload estimates with an objective psychophysiological measure. This comparison, along with analysis of the TLX, should provide a more comprehensive representation of workload during the AMBR task, incorporating subjective, psychophysiological, and predictive measures on each scenario.

Secondly, the original AMBR project lacked data on eye movements which would have helped to determine which strategies human participants used to meet task demands and how those strategies changed in different scenarios. The use of eye tracking during testing of human participants would have provided invaluable insight into the specific cognitive and oculomotor processes occurring during various levels of demand and different display types. Without data on eye movements, the crucial link between human performance model and human performance remains hidden. That is, while AMBR successfully demonstrates how well a model can simulate task behavior, it does not provide any information on the individual strategies that lead to that behavior; specifically, how tactics change and what aspects of the interface receive more attention from people in different scenarios. As impressive as it is that the models accurately predicted performance, reaction time, and subjective workload ratings, the honing of these models to take into account specific changes in strategy from scenario to scenario would be even more useful.

Comparing predictive models and descriptive workload estimates

In order to design models that provide credible predictions of cognitive workload, it is important to first validate these model predictions against objective measures of human workload. Previous research has demonstrated some of the inherent difficulties in producing reliable model predictions of workload without such validation. Schveneveldt et al. (1998) assessed the ability of a model to determine workload using information from task performance and requirements. Based on these factors a model was designed and compared with subjective workload assessment task (SWAT) ratings on three simple tasks. Although the model projections proved somewhat accurate in predicting workload ratings of human participants, the researchers concluded that these effects were well below the range of practical use and recommended that physiological workload measures be used in future modeling efforts.

More recently, efforts have been made to incorporate such physiological measures in validating model workload predictions. One such study (Son et al., 2005) used functional Near Infra Red (fNIR) technology as a means of estimating workload by measuring blood activity in the prefrontal cortex during task completion. This study compared ACT-R model predictions of workload with fNIR workload data while completing an auditory classification task at various levels of difficulty. Results indicated that ACT-R workload predictions were positively associated with blood volume activation levels, providing support for the model estimates as accurate predictions of workload experienced by human participants. The researchers acknowledge a great deal of disparity among individual physiological responses, but concluded that data from physiological observation and cognitive model prediction reveal the same general pattern. These studies demonstrate both the complexity of measuring cognitive workload and the value of comparing multiple workload estimates on a single task.

Convergent research on cognitive models and eye movements.

The link between eye movement analysis and cognitive modeling is extremely intuitive. These methodologies are often utilized in tandem to support and explain one another. Cognitive models can be used to identify a particular visual pattern as evidence of a specific cognitive strategy. Eye movement data can be used as a basis for validating cognitive models or designing others that more appropriately take these data into account. Generally speaking, in order to model human behavior with the highest possible degree of fidelity, it is important to understand precisely what the eyes are doing.

Hornof and Halverson (2003) demonstrated the applicability of eye tracking to cognitive modeling in their analysis of visual search strategies. In this study, eight models were created to simulate performance on a letter search of a static computer interface. Human performance was compared with model predictions on the variable of search time. Data from human participants fit extremely well with search time predictions for two of these models, and the other six were abandoned. The remaining models included predictions of eye movements, which were ultimately compared with the observed eye movement data. Analysis yielded several recommendations for adapting the models to fit the eye data more precisely. These included increasing foveal coverage over more than one item at a time to simulate observed peripheral vision, accounting for observed anticipatory eye movements and adapting search strategy to model an observed hierarchal approach to the search. As shown in this study, even models that fit observed data well on relevant dimensions can benefit greatly from incorporating eye movement data.

Other modeling studies have incorporated eye movement analysis with dynamic tasks, requiring completion of time-pressured objectives and other actions comparable to those of the AMBR air traffic control task. Salvucci (2005) used a driving simulation to test a model accounting for human multitasking. The model used was a version of the ACT-R cognitive architecture, modified to include a general executive capable of managing several tasks simultaneously. The model was run on the driving simulation in three different studies: An analysis of driving while operating a radio, an analysis of driving while dialing a cellular phone and an analysis of driving without any secondary task. Eye tracking allowed these researchers to compare the models simulated visual attention patterns to human eye movement patterns. Of particular interest was the amount of time spent monitoring traffic and controlling secondary devices. A comparison between the eye data and the ACT-R vision module suggested that the model simulated human visual patterns extremely well, in both monitoring of traffic and control of the radio and cellular phone. These results provided validation for ACT-R as an accurate model of visual management of a dynamic multitasking environment. These studies demonstrate the symbiotic research relationship between cognitive modeling and eye movement research.

THE CURRENT STUDY

Previous research has emphasized the importance of validating model predictions with experimental corroboration. Although subjective measures such as the NASA TLX are convenient means to this end, they lack reliability and are subject to individual biases. For this reason, the current study sought to bolster the results of the previous phase of AMBR by comparing model workload predictions with workload from a psychophysiological measure. The Index of Cognitive Activity (ICA) is a workload metric that estimates cognitive activity based on changes in pupil dilation that occur as a result of effortful processing. ICA was recorded for each participant during the AMBR simulations and used to validate the ACT-R workload predictions. Both of these measures provide moment-to-moment workload estimates as well as estimates for scenarios as a whole. In addition ICA and ACT-R predictions were compared with subjective workload ratings, performance and other facets of the AMBR task. The addition of an objective psychophysiological component to the AMBR workload comparison contributes a more precise layer of analysis in determining how human performance and model predictions differ.

As a secondary objective, the use of eye tracking to record ICA provided the opportunity to analyze participant eye movements and fixation patterns. Previous research has shown that eye tracking has the potential to provide invaluable assistance in the development and validation stages of human performance modeling. As both technologies improve, these disciplines will undoubtedly be used in conjunction with increasing frequency. The AMBR project is one such modeling endeavor that could benefit greatly from an analysis of eye movement and fixation patterns of human participants. Observed differences in performance and subjective workload ratings on the different demand levels and display conditions of the air traffic control task make it an interesting subject for an eye tracking analysis.

This report focuses on how eye tracking may provide validation of model-generated workload predictions as well as useful information regarding eye movement patterns and attention shifts during the AMBR task. By examining when participant workload increases, cognitive models may be refined to more accurately reflect human experience. By seeing what the participant sees, inferences may be made about specific strategies used to deal with the demands of different situations. These data, which were absent from the previous AMBR analysis, provide insights that allow the development of models that better predict and simulate human performance. In the following pages, eye movement and fixation patterns are examined first to cultivate a better understanding of task behavior. The workload measurements are described thereafter, followed by comparisons among the three measures.

Method

Participants

Sixteen participants, each of whom qualified for this research based on a high level of video game experience, took part in the study. Of the participants, 87.5% (n=14) were male, and 12.5% (n=2) were female. Seven participants reported an age of '21-25', five reported '26-30' and four reported '31-35.' Each participant completed all testing sessions and was compensated at the end of session three.

The Air Traffic Control Task

The present study used a modified version of an air traffic control (ATC) task to analyze eye movements and cognitive processes. The ATC interface is shown in Figure 1. The task itself, developed by MacMillan, Deutsch and Young (1997) as part of the AMBR project, requires the participant to act as an air traffic controller, assisting numerous aircraft (AC) as they enter and exit the central airspace. The AC move at a constant rate in either a horizontal or vertical direction. In this simplified version of an air traffic control environment, participants are not concerned with AC collisions. AC that appear to be on a collision course pass each other safely at different altitudes. The central airspace controlled by the participant is surrounded by four automated air traffic controllers (North, East, South and West) which are in contact with the participant throughout the task. The left side of the screen is mainly comprised of a radar screen of AC movements and locations. The right side consists of messages sent between the participant, each AC and neighboring air traffic controllers, as well as the action buttons used in response. The goal of the task is to execute a set of actions under time pressure to avoid accruing penalty points and keep AC from being delayed.

Responsibilities of the Air Traffic Controller. The first action that participants must execute during the task is to accept incoming AC as they near the yellow boundary of the central airspace. As an AC approaches, a message will appear in the incoming message window on the right side of the screen prompting the participant to accept the incoming AC. In this simplified version, the participant should always accept the AC. If the participant does not respond to the AC or responds incorrectly, the AC will stop and enter a holding pattern as soon as it reaches the yellow border of the central airspace. The freezing of the AC will be accompanied by a change in AC color, from white to red. This signals to the participant that the AC is currently on hold and actions required to free it should be taken as soon as possible. The participant will be penalized for failing to accept the AC in a timely fashion. Additional penalties will accrue for each minute that an AC remains on hold. After the AC has been accepted it will continue on its path, crossing into the central airspace.

The next action is to welcome an incoming AC. Approximately 25 seconds after an AC has been accepted, it will send a message saying 'hello' to the central air traffic controller. The participant should respond to this prompt by welcoming the AC. A penalty will be assessed for each minute that the welcome message is not sent. However, failure to welcome does not result in a holding pattern and, thus, is less important than the initial acceptance of the AC.

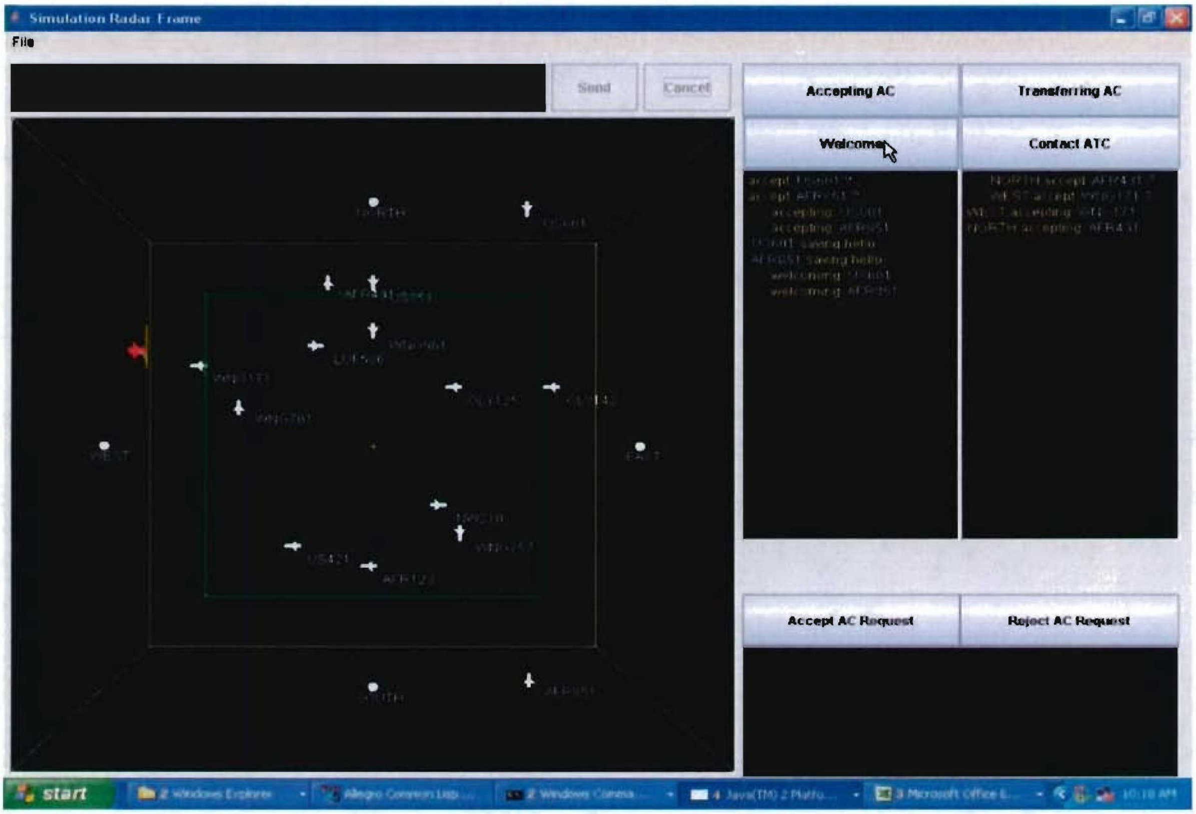


Figure 1. Air Traffic Control Display Screen

The third action is to respond to a speed request message. Over the course of each scenario, three AC will request an increase in speed. This request appears in the message module in the bottom right hand corner of the screen. Unlike all other actions in which there is only one response option, participants have to make a decision about whether to accept or reject the speed request. This decision is based on the flight path of the AC requesting a speed increase. If another AC is traveling directly in front of the requesting AC in the same direction, the speed request should be rejected. In all other situations, the request should be accepted. Failure to respond to a speed request receives a penalty for each minute the request goes unanswered. An incorrect response to a speed request also carries a penalty. As in the case of the welcome message, the speed request is considered of lower priority because failure to execute does not result in a holding pattern.

The fourth action that must be carried out is to transfer an AC that is leaving the central airspace. This is the only action that is not prompted by a message on the right side of the screen. The participant should transfer an AC as soon as the nose of the AC touches the green inner border of the airspace. From that moment, the participant has until the AC reaches the yellow outer border to transfer it to the proper adjacent controller. If the AC is not transferred in time, it will turn red and enter a holding pattern. As in the case of failure to accept an

AC, a point penalty will be assessed along with a penalty for each minute that the AC remains frozen.

The fifth and final action is to request that an outgoing AC contact the next controller. Approximately 18 seconds after an AC has been transferred, a message will appear stating that it has been accepted into a neighboring airspace. The participant must then request that the transferred AC contact the next air traffic controller. If this action is not taken by the time the AC reaches the yellow outer border, it will go into a holding pattern and the same penalties will apply. It is important to note, that for an incoming AC, there is only one compulsory action to keep the AC on its scheduled flight. An outgoing AC, on the other hand, must be both transferred and requested to contact the next controller to keep it from entering a holding pattern.

Additional Penalties. There are three additional penalties that are assessed if the aforementioned actions are not carried out accurately and efficiently:

> (1) Sometimes in the heat of the task it becomes difficult to remember which actions have already been taken. Sending a duplicate message to an AC or a neighboring air traffic controller results in a point penalty.
>
> (2) It is also important to choose the appropriate action and the appropriate AC when responding to a prompt. In addition, the participant must be careful not to respond prematurely. The action will not be accepted unless the proper prompting message or event has occurred. Sending a message that does not make sense, such as welcoming an outgoing AC or accepting an AC that has not yet requested acceptance, results in a point penalty.
>
> (3) Executing a command correctly requires that the participant first click on the appropriate action button (ACCEPT, WELCOME, TRANSFER, etc), then the AC involved and finally the SEND button. In addition, accepting, transferring and requesting contact from an AC also require the participant to click on the air traffic controller involved. The welcome and speed request actions do not require selecting a neighboring air traffic controller. Clicking on the air traffic controller unnecessarily while executing a welcome or speed request response results in a point penalty.

Display Conditions. The version of the ATC task used in this study consisted of two variations of display condition: text and color. The text display condition required participants to rely partially on the text messages on the right side of the screen to decide which actions to take. In the text display, each AC was colored white at all times unless it entered a holding pattern (at which point it turned red).

The color display provided an aid to decision making for participants. In this condition a color-coding system was used to identify which AC required attention and specifically, which type of action the AC required. As in the text condition, messages and AC positions prompted actions. However, in the color condition the AC itself changed to a different color depending on which action it required as soon as its corresponding prompt appeared. AC in need of acceptance turned green; AC in need of welcome turned blue; AC making a speed request turned magenta; AC awaiting transfer turned brown; AC awaiting a message to contact the next controller turned yellow. As soon as the appropriate action was carried out by the participant, the AC changed back to white to signal that no further actions were required for that particular AC. Eye movement data was analyzed for both text and color display scenarios. Workload analysis was completed for text scenarios only.

Levels of Demand. In both the text and color display conditions, there were three levels of demand. The number of planes requiring processing remained constant across levels. The increase in demand resulted from decreasing the length of time given to process all planes. In the lowest level of demand, scenarios lasted for 11.5 minutes, and an average of 14.9 AC were on screen at any given time (Level 1). The intermediate scenarios lasted for 9 minutes and the average number of AC on screen was 16.6 (Level 2). The highest level of demand scenarios lasted 6.5 minutes and averaged 18.5 AC on screen at a time (level 3).

Practice and Testing Scenarios. The ATC task consisted of four equivalent sets of scenarios. (A, A*, B, B*). Each contained a scenario for each of the three demand levels and for each of the two display types. The starred scenarios (A* and B*) were mirror images of the non-starred scenarios (A and B respectively). Therefore, these scenarios were judged to be comparable in difficulty. Half of the participants were trained on A and A* scenarios and tested on B and B* scenarios, while the other half were trained on a B and B* scenarios and tested on A and A* scenarios.

Description of Workload Measures

The NASA Taskload Index. The NASA Taskload Index (TLX) is a multi-dimensional rating tool designed to provide subjective assessments of operator workload in a variety of contexts. The TLX provides an overall workload score for a specific task based on ratings from six subscales: Mental Demands, Physical Demands, Temporal Demands, Performance, Effort and Frustration. It has been used in a variety of tasks ranging from flight simulations to arithmetic tasks and has been validated by Hart and Staveland (1988). In the context of past and present AMBR research, the TLX served as the tool for measuring workload on each of the air traffic control scenarios. The TLX was selected because it is convenient to administer and score and is conceptually manageable in the context of the model. The complete NASA TLX can be found in Appendix A.

The ACT-R Model. The ACT-R cognitive architecture has a long and rich history (Anderson & Lebiere, 1998). From its initial structure as a production system model, ACT-R has evolved into a hybrid architecture combining important aspects of symbolic and subsymbolic systems. One of the most important features of the architecture is its capability of simulating human performance on complex tasks by composing together many basic cognitive, perceptual and motor actions. ACT-R models can now make predictions about aspects of cognition that occur every few hundred milliseconds. ACT-R provides a description of cognition that is far above elementary brain processes but considerably below complex tasks like the AMBR Task (Lebiere, Anderson & Bothell, 2001). In its current configuration, ACT-R is highly sensitive to time pressure and high information-processing demand, making it appropriate for use in an Air Traffic Control task such as AMBR.

A model of cognitive workload under the ACT-R architecture was developed by Christian Lebiere as part of the AMBR Project (Lebiere, 2001). The model was highly successful in predicting cognitive workload, and it was demonstrated to be sensitive to level of task embedding, interaction speed, level of interface decision support, and individual differences.

The workload estimates for that model were aimed at predicting the self-reported measures of workload given by participants to the NASA TLX questionnaire. Thus, they were necessarily global estimates spanning an entire scenario because the TLX covered an entire scenario. In the research presented here, we look not only at the model's predictions on such a broad basis but also at its predictions for moment-to-moment effort and for workload-producing events that occur during the scenarios. Because an ACT-R model decomposes performance in the task in terms of each atomic cognitive, perceptual and motor step, it can generate workload predictions at any level of aggregation desired. Moreover, because ACT-R is a modular architecture (Anderson et al, 2004), it can make separate predictions for each module of the architecture, including cognitive, perceptual and motor workload. Finally, ACT-R is not a normative model of cognition but can instead capture individual differences through knowledge and parameter variations, which can provide a measure of individual workload (Rehling et al, 2004).

The model developed by Lebiere (2001) to predict workload in the AMBR task was adapted slightly in order to ensure that it was constrained for performing the same actions in the same time frame as participants, a techniquc known as model tracing. Model tracing consists of forcing the model to follow an execution path that is closest to that of the subject. The goal is to keep the context similar for subject and model throughout the simulation run to be able to make meaningful comparisons for the entire data set instead of just the part until which they diverge and make comparisons meaningless. If model and subject were in different situations, comparing their workload for that particular time interval would be meaningless. That means:

- Identifying the decision points in the model, which provide for the possibility of multiple future paths. In this case, those consisted of conflict resolution sets with more than one production, memory retrievals that matched more than one chunk, and perceptual events with multiple possible outcomes.

- Identifying the future events in the trace of the subject run that determine the model path. In this work, those were the external actions of the participants that could be unambiguously attributed to a particular state, e.g. selecting buttons or objects on the screen.

- Formulate a method by which to choose at the model decision points based upon future subject events. The algorithm used in the work presented here was a one-to-one correspondence between subject event and model path.

The only alterations made from the original ACT-R model of the AMBR task were made in order to accommodate the model tracing procedure. The current version of the model is described in Appendix B, with particular attention devoted to describing any alterations that were made from the original model. Note that the original model was developed in order to perform two versions of the AMBR task, a color version, and a text version. The current effort only focused on the text version of the task, and the following discussion only consists of those aspects of the model relevant to the text version[1].

The Index of Cognitive Activity. The Index of Cognitive Activity (ICA) is a patented psychophysiological measure that estimates cognitive workload based on changes in pupil dilation (Marshall, 2000). The ICA has been used in a number of applications, including problem solving, decision making, and augmented cognition (Marshall, Pleydell-Pierce, & Dickson, 2003; Marshall, 2005; Marshall, in press).

The ICA is based on the well-known fact that the pupil dilates during effortful cognitive processing (Loewenfeld, 1993). The most common technique used to assess pupil dilation has been the task-evoked pupillary response, developed by Jackson Beatty and his colleagues (Beatty & Lucerno-Waggoner, 2002). Although the ICA technique and the task-evoked pupillary response are based on different methods of analysis, they have been shown to produce similar results when used in the standard digit span task (Marshall, Davis, & Knust, under review). This task was originally used by Beatty to demonstrate that the pupil changes systematically as the task dimensions change. Both the ICA and the task-evoked pupillary response produce statistically significant linear trends such that pupil activity increases as the digit span to be recalled increases. The advantage of the ICA over the task-evoked pupillary response is that the ICA can be applied meaningfully across a scenario of complex events for a single individual, providing event-based as well as time-based estimates of cognitive effort.

The nearly continuous recording of pupil size is a signal that can be processed like any other signal. The ICA is calculated from high frequency components of this signal as an individual performs a specific task. It is a measure of relative change that reflects the number of times each second that abrupt increases in the amplitude of the pupil signal occur. High ICA values reflect increases in the number of bursts of dilation by the pupil and correspond to considerable mental effort. Low ICA, on the other hand, reflects a relatively calm pupil and little mental effort. ICA has proved effective as a measure of workload on a variety of tasks and is capable of distinguishing between cognitive states ranging from focused attention to boredom and fatigue (Marshall, 2005).

The calculation of ICA used in the current study was based on a pupil signal recorded at 250 Hz. The eye-tracking system used in the study was the EyeLink II (from SR Research, Ltd.), a binocular system that records both pupil size and horizontal and vertical point of gaze for each eye every 4 msec. Prior to ICA computation, the point of gaze data were analyzed to determine the times at which unusually rapid eye movements or unusually large saccades occurred. The pupil measurements corresponding to these unusual movements were then eliminated from the pupil signal by linear interpolation. Full blinks and partial blinks were also eliminated, with their corresponding pupil values replaced by linear interpolation. Wavelet analysis was then applied to the resulting pupil signal, and a statistical threshold was used to determine which wavelet coefficients were unusually large. The frequency and location in time of these large coefficients forms the basis of the ICA, as described in Marshall (in press). ICA estimates were computed both overall and second by second for each scenario run. Finally, to make comparisons with the ACT-R results easier, the ICA estimates were transformed into a range of 0-1 through the hyperbolic tangent.

As in the case of ACT-R workload predictions, ICA is sensitive to stimulus complexity, making it a viable option for estimating workload on the air traffic control task. In addition, ICA can provide both sub-second estimates of workload and global estimates aggregated over entire scenarios. When compared with ACT-R estimates, ICA aids in determining how closely the model predicts workload as reflected by an objective psychophysiological measure.

Procedure

Eye-Tracking Procedure. The data reported here were collected using the EyeLink II Eye-Tracking System from SR Research, Ltd., with binocular tracking at a sampling rate of 250 Hz. The EyeLink II System consists of small video cameras mounted on a lightweight headband. Two cameras record eye data while a third camera records the position of the head, allowing a reasonable range of movement. The system offers gaze position error less than .05°.

The ATC task screen was divided into 17 regions corresponding to the different sections of the screen. These regions cover each of the action buttons and message windows as well as key features of the radar screen of AC positions. The display screen and all regions can be seen in Appendix B. The areas of the screen examined most carefully were the regions comprising the radar screen and those comprising the message windows (see Figure 2). Eye data were analyzed to determine the percentage of total viewing time spent in each region and the number of transitions between regions.

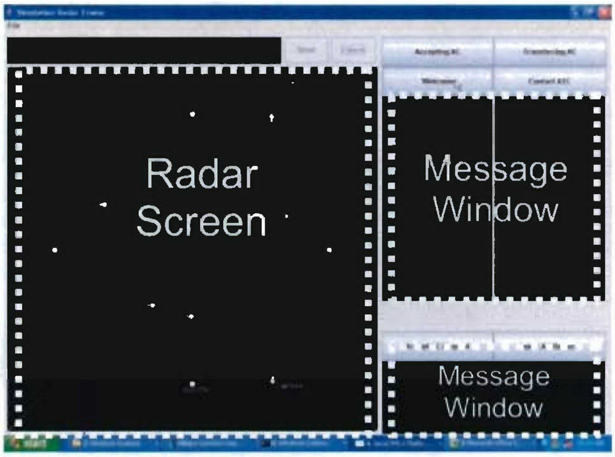


Figure 2: Analysis focused heavily on these areas

Experimental Procedure. Participation in the study consisted of three experimental sessions over the course of one week, with a day off in between each session. To begin the first session, participants completed a questionnaire regarding previous video game experience and were introduced to the procedures and equipment used in eye tracking research. Once participants were comfortable with these aspects of the project, the experimenter proceeded to explain the ATC task. The first session consisted entirely of training and practice. The experimenter explained each aspect of the rules and penalty point system, providing demonstrations as needed. Participants were then given the opportunity to practice independently, both with and without guidance from the experimenter. The participant finished session one by completing two scenarios while eye movements were measured.

Session two involved the completion of six scenarios by each participant without the assistance of the experimenter. The order of display condition was counterbalanced such that odd-numbered participants began with three text display condition scenarios and then, after a short break, completed three color display condition scenarios. Even-numbered participants began with three color condition scenarios and then, after a short break, completed three text condition scenarios. Participants always completed demand Level 1 scenarios first, Level 2 scenarios second, and Level 3 scenarios last (which was the order used by the original AMBR study). After completion of each scenario, participants completed TLX workload ratings. The experimenter recorded eye movements and pupil dilation on all six scenarios for session two. Although eye tracking was employed in these first two sessions, this was primarily to allow the participant to grow accustomed to performing the task while wearing the eye tracking headset. Data from these first sessions are not included in the analysis.

Session three proceeded in much the same way as session two. Participants completed six scenarios with a short break in the middle. The specific scenarios completed were slightly different than in session 2, but the order of display condition and demand level remained the same. TLX workload ratings were again taken after each scenario. After completion of the final scenario, participants completed a follow-up questionnaire, received a debriefing, and were dismissed. Eye movements and task performance data from this third session, after participants had received extensive training and practice, are the focus of the current analysis.

Workload Analysis Procedure. A data file recording gaze position at 250 Hz was used to analyze participant gaze and transitions between regions. Each completed scenario corresponded to a separate eye data file. These files were then adjusted to account for slight shifts in the equipment so that the data files reflected as exactly as possible where participants were looking at all points during the tasks. This allowed participant data to be aggregated across subjects or scenarios or analyzed on a single subject basis.

On screen stimuli, aircraft movements, and participant actions were summarized in log files to be compared with workload data. These files organized all events and actions chronologically along with aircraft coordinates and programming terminology to provide an individualized textual representation of each scenario. The large size and unwieldy format of these files necessitated a further refinement to exclude irrelevant information and organize all pertinent events and actions on a second by second basis.

The resulting condensed data files contained 13 variables calculated on a second-by-second basis. Eight variables recorded the various dimensions of the stimuli on the display. The first variable provided information on how many AC were moving during each second of each scenario. Five variables detailed the occurrence of each of the five stimuli requiring action (e.g. plane requests acceptance, plane crosses border, etc.) with an additional variable capturing the overall count of stimuli for each second. A final stimulus variable provided information on when an aircraft entered a holding pattern (the most critical error in the simulation).

Two response variables were coded. The first specified the time at which a response option was chosen (e.g. 'transfer aircraft,' 'send welcome message,' etc) and the second specified the time at which the send button was clicked.

Two final variables captured global aspects of the simulation. The variable *Taskload* was calculated by summing the number of stimuli awaiting response. For example, if there were three stimuli to which the participant had not yet responded, *Taskload* would be three. As soon as responses were made to address those three stimuli, *Taskload* would return to zero until another stimulus event occurred. *Taskload* was used extensively in the analyses and was considered especially relevant because it revealed periods when participants were overloaded with actions requiring completion. The final variable was a total *activity* score. This variable summed all stimuli, response option choices, responses and the *Taskload* for each second. This total *activity* score served as a composite of on-screen action as a way of revealing the time periods during which the most workload was taking place. Presumably, these periods of time, marked by increased stimulus occurrence, button clicks and accumulating *Taskload,* should be associated with increased ACT-R and ICA estimates of workload. Enhanced AMBR log files containing these variables were created for each scenario of the final day of testing, yielding a total of six files for each participant.

The workload estimates reported here were gathered from three sources:

(1) The TLX was completed at the end of each scenario by each participant. All scales were averaged to provide a total TLX score.

(2) ACT-R workload estimates were calculated every 50 milliseconds basis and aggregated across five second intervals. These intervals were used to calculate total task averages and for comparison with five second ICA intervals.

(3) ICA was recorded at 250 Hz and summarized on a second-by-second scale, as well as for entire scenarios. The second-by-second demarcation was further aggregated over five and ten second intervals. The ICA provides workload estimates for the left and right eye separately as well as an averaged total of both eyes. The right eye was found to be the better estimator for the current study and was used in the analyses here.

All workload measures were compared with each other and with data from the condensed AMBR logs. This study provided the opportunity to analyze specifically which elements of the task were influencing subjective, psychophysiological, and predictive measures of workload.

RESULTS I: HUMAN SUBJECT COMPARISONS

Before approaching the eye data, human subject variables were analyzed for comparison with the previous AMBR sample. The three dependant measures examined were performance, TLX workload ratings, and reaction time. Performance data consisted of the averaged score in penalty points on each scenario. Workload ratings by each participant were averaged to create an aggregate workload rating for each task. Reaction time was determined by calculating the average difference between a stimulus occurring on screen and a proper action taken in response. These dependent measures were the same three measures used in previous AMBR research.

Results from performance data indicated that participants accrued more penalty points in text display scenarios than in the color display scenarios, *F* (1,82)=21.14, *p*<.001. Performance also suffered as overall task demand level increased, *F* (2,81)=3.26, *p*<.05, and the interaction between display condition and demand level was significant, *F* (2,81)=4.89, *p*<.01. Because participants generally performed near perfection on color display scenarios at all three levels, the effect of scenario demand was only significant in text display scenarios, *F* (2,81)=5.21, *p*<.01. This performance data corroborates the findings of previous AMBR research; participants perform more poorly on the text scenarios, especially as task demand level increases.

Results from reaction time data suggested that participants reacted to on-screen events more quickly in the color display condition and on lower demand scenarios. An analysis of variance yielded significant main effects for display condition, *F* (1,82)=55.56, *p*<.001, and demand level, *F* (2,81)=13.42, *p*<.001, and a significant display condition by demand level interaction, *F* (2,81)=8.05, *p*<.001. Unlike the performance effects, these reaction time effects were significant independently for both color and text scenarios. Participants reacted more slowly on text scenarios than on color, and this effect was exacerbated by increase in demand level. These results concur with those of the previous AMBR sample.

Participant TLX ratings on each of the six scales of the TLX were averaged to create a total TLX rating for each scenario. Analysis of this total TLX score yielded significant main effects for display condition, *F* (1,82)=21.26, *p*<.001, and level of demand, *F* (2,81)=3.19, p<.05. The interaction was not significant. As in the case of performance effects, TLX ratings did not differ significantly across levels of demand on color display scenarios taken independently. In other words, the main effect for level of demand was primarily caused by differences in TLX ratings in the text scenarios. These results are in agreement with data from previous AMBR subjects with one exception; the previous study demonstrated a significant main effect of demand on color display scenarios independently, and the current study did not.

RESULTS II: PARTICIPANT VIEWING PATTERNS

Display Condition Differences

As shown in Figure 3, participants exhibited significantly divergent viewing patterns on text display and color display scenarios. The major difference was the amount of time spent viewing the radar screen on the left side of the display. This area contained a radar plot of the central sector that the participant controlled, as well as part of the four neighboring air traffic controllers' sectors. In the text display condition, participants spent 57.6% of viewing time monitoring aircraft movement in this part of the screen. During color display scenarios, participants spent 72.0% of total scenario time viewing this area. This result was accounted for by differences in viewing percentages in the message windows. This area contained all windows on the right side of the screen which displayed incoming and outgoing messages. During the text display condition, participants spent 38.4% of total viewing time in this part of the screen, compared with only 22.9% in color display scenarios.

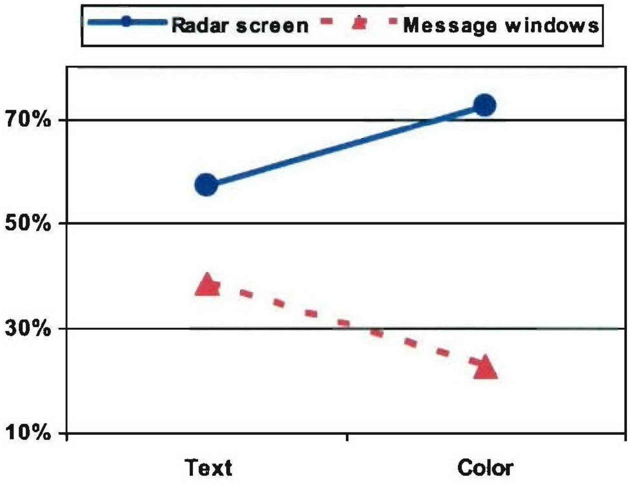


Figure 3. Viewing Percentages on text and color scenarios

Further analysis of the viewing statistics revealed specifically which regions received the most attention in the different display conditions. The *outgoing AC messages* region exhibited the greatest increase in viewing time in text scenarios, *F* (1,82)=107.62, *p*<.001, followed by the *incoming AC messages* region, *F* (1,82)=67.14, *p*<.001. The *inner square* region demonstrated the greatest increase in color display viewing percentage compared to text, *F* (1,82)=52.08, *p*<.001. Other regions influenced to a lesser extent by display type included *west inner border* and *north exterior,* which received more attention in color display scenarios, and *speed request messages* and *outgoing response options*, which received more attention in text scenarios.

Demand Level Differences

Level of demand impacted participant viewing patterns as well. These differences were most evident when text and color were separated. In text scenarios, percentage of time viewing the radar screen decreased as demand level increased. Participants spent 60.0% of scenario time in this area during level 1, compared with 57.7% in level 2 and 54.8% in level 3. As attention to this area decreased, attention to the message windows increased; during level 1 24.0% of time was spent in the message windows compared with 25.9% in level 2 and 27.7% in level 3. In color scenarios, there was not a clear change in percentage of time spent on the radar screen at different levels of demand. All levels of color scenarios averaged near the grand mean of 72.0%. However, there was a noticeable trend regarding attention to the message windows. As level of demand increased, percentage of time viewing the message windows decreased. Participants spent 13.5% of scenario time in this area during level 1, compared with 10.2% in level 2 and 8.8% in level 3.

Further analysis of viewing statistics revealed which regions received the most viewing attention during different levels of demand. In text conditions, the *speed request messages* region exhibited the greatest increase in viewing time as demand increased, F (2,39)=21.26, p<.001, followed by the *outgoing AC messages* region, F (2,39)=5.47, p<.008. The *inner square* region demonstrated the greatest decrease in viewing percentage as demand increased, F (2,39)=15.01, p<.001. Another region influenced by demand level was *outgoing response options,* which received more attention as demand increased.

In color scenarios, the *speed request messages* region exhibited the greatest decrease in viewing time as demand increased, F (2,39)=4.65, p<.015, followed by the *outgoing AC messages* region, F (2,39)=3.25, p<.05. The *west exterior* region demonstrated the greatest increase in viewing percentage as demand increased, F (2,39)=7.04, p<.002. No other regions demonstrated significant differences across demand levels.

Display Condition and Demand Level Interactions

The combined main effects of display condition and level of demand yielded interactions meriting further analysis. Display type and level of demand interacted to influence percentage of time spent viewing the radar screen, F (2,81)=7.032, p<.011. During text display, increases in demand level caused decreases in viewing time in this area of the screen. Contrarily, during color display, increases in demand caused increases in viewing the radar screen. Specific regions affected by interactions causing a similar trend included the *inner square* and *north exterior* regions.

Display type and demand level interacted to facilitate an opposite pattern in percentage of time spent viewing the message windows, F (2,81)=35.78, p<.001. During the text display conditions, as demand increased participants spent more time viewing the message windows. In color display conditions, on the other hand, increases in demand were met with decreases in viewing time for the message windows. Specific regions affected by interactions causing a similar trend included the *incoming AC messages* and *outgoing AC messages* regions.

Participant Transition Patterns

For the purposes of this analysis, a transition was defined as any movement of visual fixation from one region of the screen to another. Of particular interest were transitions between the radar screen and the message windows. An analysis of the average number of transitions between these two screen areas per second yielded significant differences, F (1,82)=11.95, p<.004. During text display scenarios, participants shifted gaze between the radar screen and the message windows significantly more than in color scenarios. In all scenarios, transitions most frequently involved the *inner square* region, followed by the *incoming AC messages* and the *outgoing AC messages* regions. Level of demand significantly affected transitions per second between the radar screen and message windows, F (2,81)=15.58, p<.001. As demand level increased, transitions per second between these regions decreased.

Analysis of the human performance data alongside the transition data yielded some interesting associations. On text scenarios, transitions per second between message windows and radar screen correlated negatively with score when controlling for demand level, r (39)=-0.31, p<.05. This suggests that poor performance was associated with a decrease in transitions, regardless of demand level. This trend was not evident in the color scenarios, probably due to the lack of variance in scores on color scenarios. The color scenarios did suggest an association between number of transitions and certain facets of the TLX scale. As number of transitions per second increased, participants rated the scenarios as being more mentally and temporally demanding, r (39)=0.35, p<.03 and r (39)=0.39, p<.02 respectively.

RESULTS III: WORKLOAD COMPARISONS

The analysis of viewing percentages and transition patterns between different screen areas provided valuable insight into the strategies human participants performed in the different display conditions and demand levels of the task. This information, together with performance measures and questionnaire responses provides a fairly comprehensive framework for understanding and modeling the AMBR task. The rest of this report seeks to address another critique of the original AMBR study, namely the subjectivity of the TLX. The major focus of the current study was to provide a more robust measure of workload with which to compare the model predictions of workload. The use of ICA, a metric based on fluctuations in pupil diameter, allowed comparison of model predictions with a physiological measure of cognitive activity.

Originally we had planned to examine the ICA in a format similar to the viewing and transition analysis, looking primarily at differences between display conditions and levels of demand. However, after extensive analysis into the ICA patterns during the color display condition, it became clear that our estimates of workload did not reveal a predictable pattern of cognitive activity. This is not to say that the ICA did not effectively measure the cognitive activity of each participant. Previous research has demonstrated the validity of ICA as a measure of effortful cognitive processing (Marshall, 2005). The reason that we were unable to elucidate any specific patterns of cognitive activity in the color scenarios seems to have been related to the lack of effortful cognitive processing required to meet the demands of these tasks. Figure 4 shows a visual depiction of a portion of a color scenario. The colored areas indicate the occurrence of bursts of cognitive activity based on fluctuations in pupil diameter. The large amount of pupil activity in the center of the radar plot, where very few important stimuli occur suggests that the cognitive activity is not in response to on-screen events. This can be contrasted with figure 5 which is a visual depiction of ICA on a text scenario. Notice that the ICA is primarily distributed on the borders of the central sector where most of the necessary action prompts take place.

We offer the following explanation for these results. In the color scenarios, after participants had learned the specific actions required for each color change in stimulus, very little meaningful cognitive processing occurred. There was no longer any need to cognitively process the radar screen for planes needing assistance, nor was there any reason to analyze the contents of the message windows. Participants relaxed their focus and waited for the changing colors to capture their attention. Then they preformed the set of actions required to satisfy the aircraft demands. Questionnaire ratings and scores from these scenarios revealed how effortless participants found these color scenarios to be at all levels of demand. Participants averaged .54 errors on color scenarios compared with an average of 9.55 on text scenarios. On the follow-up questionnaire, participants rated the difficulty of color scenarios at 1.63 on a 10-point scale, whereas ratings for text scenarios averaged 6.69. TLX ratings also reflected this trend, as demand level did not significantly impact subjective workload ratings, *F* (2,39) = .12, *p*<.9. On an anecdotal level, we note that two participants fell asleep during the color scenarios and others appeared bored. Several offered comments about how easy this task was.

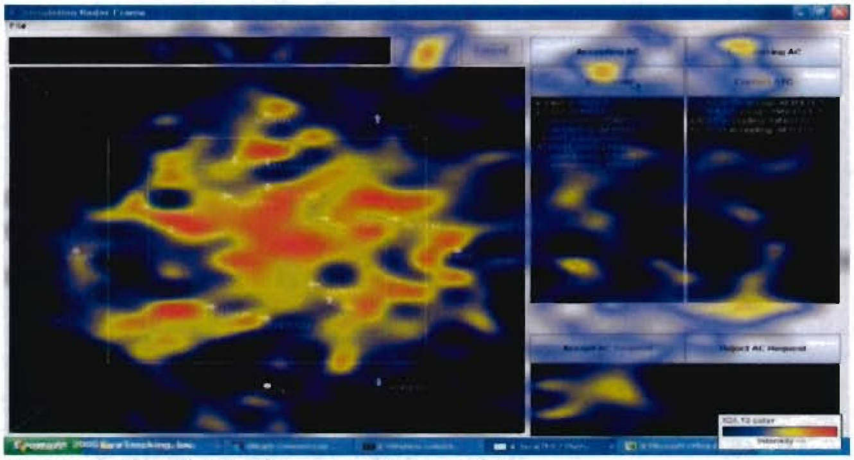


Figure 4. ICA distribution on a color scenario

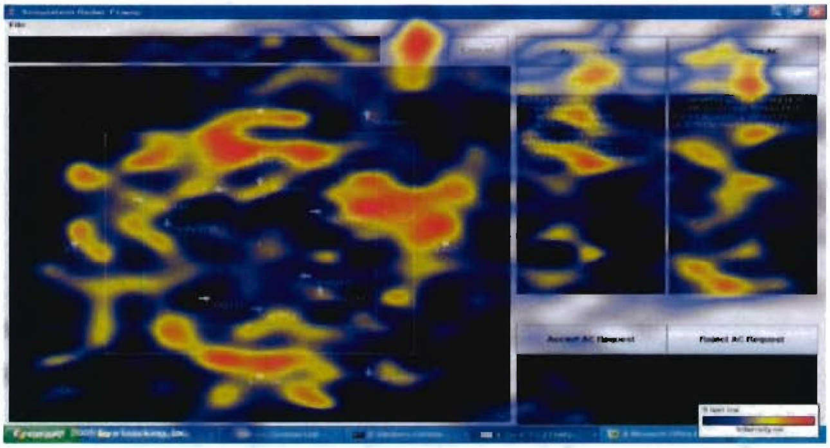


Figure 5. ICA distribution on a text scenario

Consequently, the analyses in the remainder of this report focused on the more difficult text version of the task. The color scenarios were excluded from the comparisons of workload estimates.

General results for three measures of workload

Descriptive statistics for all three workload measures are presented in Table 1.

Table 1. Descriptive statistics for workload measures on the text scenarios

	n	Mean	Standard Deviation	Minimum Value	Maximum Value	Range
TLX	48	3.78	1.97	.50	7.75	7.25
ACT-R	48	0.55	0.04	0.47	0,64	0.17
ICA	48	0.40	0.07	0.24	0.56	0.32

As predicted, all three workload measures revealed a significant linear trend of workload across the three demand levels, with higher workload associated with higher demand level for all measures. Statistical tests of linear trend, based on repeated measures analyses of variance, were *F*(1,15)=82.45, *p*=.000; *F*(1,15)=14.48, *p*=.002; and *F*(1,15)=40.52, *p*=.000 respectively for ACT-R, ICA, and TLX measures. The workload means for each demand level together with means for score and *Taskload* are displayed in Table 2.

In addition to workload measures, score and *Taskload* also showed a significant linear trend across the three demand levels of the scenarios, with *F*(1,15)=22.66, *p*=000 and *F*(1,15)=77.03, *p*=.000 respectively. The sensitivity of these two variables to all three levels of demand suggests that demand level was successful in fostering a higher level of difficulty of the scenario. A higher score indicates more errors and a higher *Taskload* reflects an inability to complete tasks in a timely fashion.

Table 2. Means for each workload measure and two other variables at each level

	TLX	ACT-R	ICA	Score	*Taskload*
Level 1	2.67	.512	0.391	2.31	0.49
Level 2	3.80	.546	0.393	11.66	1.26
Level 3	4.88	.586	0.414	15.38	1.79

Cross Validation of Workload Measures: Overall

We first examined sets of correlations among the variables of interest. Table 3 shows nine different sets, each spanning all three levels of task difficulty. None of the 27 correlations of Table 3 reach the level of statistical significance. Surprisingly, neither ACT-R nor ICA correlated highly with NASA TLX on any level of difficulty. On the contrary, these correlations suggest an inverse relationship between TLX and the other two workload estimates.

Part of the problem may be the low power associated with the tests of correlation. For each computation, we have only 16 pairs of data. However, the results are surprising, given that each of the five variables independently showed a significant linear trend across the three levels of difficulty. The correlations suggest that each of the variables may be detecting unique aspects of the difficulty imposed by the task.

Table 3. Intercorrelations among workload measures

	ACT-R1	ACT-R2	ACT-R3	ICA1	ICA2	ICA3	TLX1	TLX2	TLX3
ACT-R1				-0.40			-0.13		
ACT-R2					0.32			-0.34	
ACT-R3						0.12			-0.39
ICA1	-0.40						-0.10		
ICA2		0.32						-0.28	
ICA3			0.12						-0.34
TLX1	-0.13			-0.10					
TLX2		-0.34			-0.28				
TLX3			-0.39			-0.34			
Score1	0.08			0.14			0.17		
Score2		0.15			0.15			0.43	
Score3			-0.01			-0.04			0.35
Taskload1	0.16			-0.02			0.38		
Taskload2		0.02			0.36			0.42	
Taskload3			0.29			0.20			0.47

Exploration of 5-Second Intervals of ACT-R and ICA

Although both ACT-R and ICA can be examined at a smaller scale, we elected to examine them at 5-second intervals throughout each scenario. This time was selected to minimize noise that might otherwise confound the results.

A key analysis was the examination of the range and distribution of workload values that emerged in the study. The three scenarios had 138, 108, and 78 intervals, with each interval spanning 5 seconds. Recall that scenarios under the three levels differed only by time pressure. Each one had the same number of aircraft that required attention, and each had the same number of participant responses that must be made. An important feature of this task is that *no new decisions or cognitive action was required* in the more difficult levels than was required on the easiest level. The basic task of identifying an aircraft correctly and making the appropriate response occurred repeatedly without variation across the three levels of difficulty. What then causes one level to be more difficult than any other? The answer is simply time. The same number of events occur in all three levels of difficulty, but they occur in closer proximity in the most difficult level.

Consequently, we raise the hypothesis that workload for this task is simply due to time pressure. The same number of cognitive actions must be made in all three levels, but they must be carried out in increasingly shorter time as the difficulty level increases. If that is the case, then we ought to observe approximately the same degree of peak workload (i.e., the amount of workload measured for any 5-sec interval) across the three levels.

A simple test confirms our hypothesis. To make the test, we first created frequency distributions for each individual on each text scenario on day 2. That is, we took the final day of testing and looked at text scenarios of all three difficulty levels. For both ACT-R and ICA workload estimates, we defined 12 categories, as shown in Tables 4 and 5. The lowest levels of workload, corresponding essentially to no effort, were eliminated from both workload measures. For ICA, values below .2 were ignored, and for ACT-R, values below .34 were similarly ignored. (The ACT-R value of .34 corresponds to minimal workload and is the lowest value recorded by the model.)

The distributions of Tables 4 and 5 are informative. Consider the ICA values of Table 4. The marginal sums of 1902, 1513, and 1152 account for 86%, 88%, and 92% of all 5-sec intervals across all levels for all participants. As would be expected, the more difficult the level, the fewer the number of low workload values. The most striking feature of Table 4 is the number of intervals that fall into the highest three workload categories of >.650, .611-.650, and .571-.610. If Level 3 were truly more cognitively effortful than Levels 1 and 2, we would expect to see a larger proportion of intervals in these highest workload categories, because participants should be exerting greater cognitive effort to cause higher values if ICA. This is not the case. As Table 4b illustrates, the proportions for each level falling into each bin is roughly equal. A simple χ^2 test of proportions shows that the proportions do not vary across the levels, with $\chi^2(22)=16.03$, which is far below the 33.92 value required for significance.

Tables 4a and 4b show the aggregate data for all 16 participants. Tests were conducted on each participant's data, and all show the same pattern. No χ^2 test had a significant difference in proportions across the three levels. Moreover, for 11 of the 16 participants, the very highest 5-sec interval recorded was *not* in Level 3 but in Levels 1 or 2.

Table 4a. Frequency distribution for ICA values across three levels of difficulty.

	Level1	Level2	Level3	Total
>.650	89	75	70	**234**
.611-.650	75	48	47	**170**
.571-.610	97	74	59	**230**
.531-.570	135	103	75	**313**
.491-.530	170	118	111	**399**
.451-.490	188	156	117	**461**
.411-.450	204	174	107	**485**
.371-.410	219	185	142	**546**
.331-.370	225	169	138	**532**
.291-.330	204	163	128	**495**
.251-.290	159	130	78	**367**
.211-.250	137	118	80	**335**
	1902	**1513**	**1152**	**4567**

Table 4b. Proportions of intervals falling into each category for each level

	Level1	Level2	Level3
>.650	0.05	0.05	0.06
.611-.650	0.04	0.03	0.04
.571-.610	0.05	0.05	0.05
.531-.570	0.07	0.07	0.07
.491-.530	0.09	0.08	0.10
.451-.490	0.10	0.10	0.10
.411-.450	0.11	0.12	0.09
.371-.410	0.12	0.12	0.12
.331-.370	0.12	0.11	0.12
.291-.330	0.11	0.11	0.11
.251-.290	0.08	0.09	0.07
.211-.250	0.07	0.08	0.07

A similar analysis was made on the ACT-R workload estimates. Tables 5a and 5b provide the frequency distribution and proportions. The marginal sums of 1055, 976, and 869 account for 46%, 56%, and 70% of all 5-sec intervals across all levels for all participants. Thus, there are many more very low ACT-R intervals than ICA intervals. Recall that .34 is the lowest value estimated for ACT-R, so the cutoff levels for the two workload measures are quite similar.

Table 5a. Frequency distribution for ACT-R values across three levels of difficulty.

	Level1	Level2	Level3	Total
>.960	92	63	32	**187**
.911-.960	20	34	22	**76**
.861-.910	81	84	72	**237**
.811-.860	81	69	70	**220**
.761-.810	73	83	72	**228**
.711-.760	101	121	96	**318**
.661-.710	102	107	101	**310**
.611-.660	123	107	121	**351**
.561-.610	114	100	90	**304**
.511-.560	96	74	81	**251**
.461-.510	91	70	54	**215**
.411-.460	81	64	58	**203**
	1055	**976**	**869**	**2900**

Table 5b shows the proportions of 5-sec intervals falling into each workload category. As with the ICA, the proportions across the levels for each category are very similar. The biggest discrepancy comes in the highest workload category (>,960) in which there are many *fewer* such intervals at the most difficult level.

Table 5b. Proportions of intervals falling into each category for each level

	Level1	Level2	Level3
>.650	0.09	0.06	0.04
.611-.650	0.02	0.03	0.03
.571-.610	0.08	0.09	0.08
.531-.570	0.08	0.07	0.08
.491-.530	0.07	0.09	0.08
.451-.490	0.10	0.12	0.11
.411-.450	0.10	0.11	0.12
.371-.410	0.12	0.11	0.14
.331-.370	0.11	0.10	0.10
.291-.330	0.09	0.08	0.09
.251-.290	0.09	0.07	0.06
.211-.250	0.08	0.07	0.07

As with the ICA estimates, a χ^2 test of proportions was used to analyze these data, but here the results were significant, with $\chi^2(22)=42.94$, $p<.05$. However, most of the χ^2 contribution comes from the first cell of Level 3 and shows that there are significantly fewer intervals in this cell than would be expected. Individual χ^2 tests for all participants were uniformly non-significant.

Thus, both ICA and ACT-R workload estimates show the same pattern of values. There is no evidence of greater cognitive effort in the more difficult levels; participants do the same tasks and apparently expend the same effort whether we estimate it by psychophysiological sensors or through cognitive models.

Why then do the two workload measures not agree? We believe the answer to be that it is very difficult to align the exact instance that the cognitive effort is detected by these two metrics. Because the events within the scenarios are sometimes overlapping, it is impossible to tell which event has received the participant's attention and consequently which one is driving the cognition at that moment. This issue will be revisited in Discussion II below.

DISCUSSION I: EYE-MOVEMENT ANALYSIS

Lessons from Eye Movements

In designing models that accurately simulate and predict human performance, it is important to take into account situational variability that may influence behavior. Human beings are capable of adapting rapidly to changes in their environment with updated strategies that deal more effectively with obstacles. This ability to shift strategies, often unpredictably, poses a major challenge to the precision of human performance models. For this reason, it is important to analyze specifically how and when people change their approach to solving a particular problem. In the current study, we were able to analyze human strategy changes on the air traffic control task as they related to differences in color and text display condition and differences in task demand. Through the use of eye tracking, it was possible to objectively examine the manner in which participants visually interacted with and responded to the changing characteristics of the task.

Analysis of differences between color and text display conditions in eye movements proved extremely revealing. These scenarios were indistinguishable in terms of AC movement, number of actions to perform and duration of each scenario. The only difference was the addition of color-coding to AC in the color display condition. Substantial discrepancies in task performance, reaction time, and TLX workload ratings between these two conditions suggested that this color coding had a significant effect on the way that participants interacted with the otherwise identical scenarios. By examining eye movements, we were able to determine whether or not the strategies employed by participants differed.

Differences in percentage of time viewing particular areas of the screen are indicative of divergent strategies. Participants in the color display scenarios spent significantly less time viewing the message windows than did participants in the text display scenarios. Although the majority of time in both conditions was spent viewing the radar screen, this dip in message viewing time for color scenarios suggests that the area of focus for identification of required actions had shifted form the messages to the color coded aircraft. In particular, the *inner square* region received a significant boost in viewing time at the expense of time spent in the *outgoing AC messages* region. Another aspect of the strategy featured in the color display scenarios is revealed by the transition data. Participants shifted gaze less frequently between messages and the radar screen in the color condition. Without the burden of reliance on text messages, participants were able to limit the amount of switching between message viewing and radar screen viewing.

The interaction between color and demand level demonstrated how participants dealt with increases in task demand in different display conditions. As demand increased, text and color viewing patterns reacted inversely. In the text display conditions, increases in demand were met with increases in viewing the message windows and decreases in viewing the radar screen. Just the opposite occurred in the color display condition. Demand increases were met with decreased viewing time in the message windows and increased viewing time in the radar screen. Display condition and level of demand did not interact to influence transition rates. The effect of level on transition rates was consistent for both display conditions; as demand increased, transitions per second decreased. Based on these analyses of viewing patterns and transition rates, it is possible to generate a cohesive description of general strategies on the different scenarios.

Two Distinct Viewing Strategies

Participants in color display scenarios tended to focus their attention almost entirely on the radar screen. The color cues allowed important stimuli to be detected without consulting the message windows. For this reason, there was significantly less shifting of attention between the message windows and the radar screen as shown by the diminished rate of transition. As the level or demand increased in these color scenarios, the amount of time spent looking at the message windows decreased even further. Because of the added workload in these more difficult scenarios, participants compensated by further restricting the amount of time spent monitoring the text. Transition rates also decreased, suggesting more careful focus on one particular area of the screen as opposed to frequent shifts of focus. Essentially, when demand levels necessitated a faster rate of response, participants in color display scenarios focused their attention heavily on the radar screen. This strategy proved particularly effective, as reflected by high levels of performance and low TLX workload ratings. The color display viewing strategy is illustrated in Figure 6, which graphically renders a participant's gaze as distributed over an entire scenario.

In text display scenarios, participants adopted a very different strategy to handle the demands of the task. As in the color display scenarios, the radar screen received the most attention. Monitoring AC in this sector of the screen was most important in all scenarios. However, in the text display scenarios, participants also focused heavily on the message windows. Because important stimuli were not cued by color-coding, careful scrutiny of the message windows became essential to the task. Thus, participants allocated a larger portion of their time to the message windows than did participants in the color display condition.

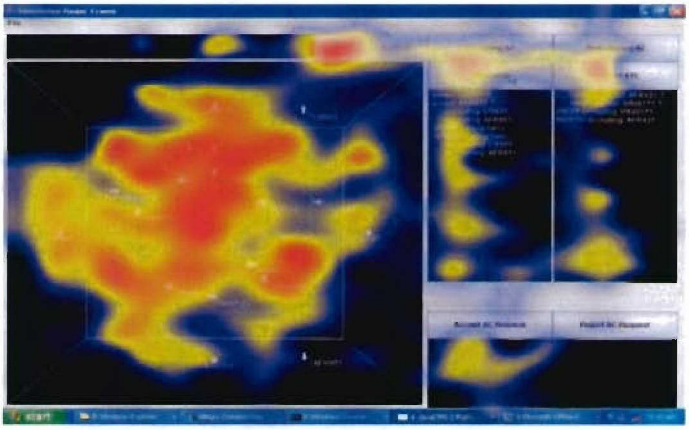


Figure 6. Example Gazespot of color scenario

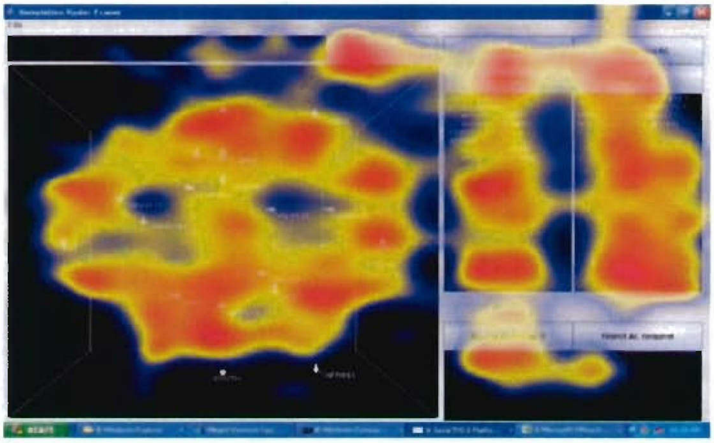


Figure 7. Example Gazespot of text scenario

In addition, participants gaze shifted more frequently between the message windows and the radar screen. The chosen strategy seemed to involve frequent movements back and forth as opposed to alternating longer periods of time monitoring each area. As level of demand increased in the text scenarios, participants assumed a strategy opposite to the strategy during color scenarios. Participants responded to increases in demand level with increased time spent viewing the message windows, instead of the radar screen. In text display scenarios participants could not simply shift their attention to the radar screen to deal with increased workload. Instead they spent additional time looking at the message windows, waiting for new messages to appear and searching for old ones that they may have missed. Transition rates decreased at these higher levels of demand, suggesting that participants took more time in scrutinizing each area before shifting attention. It is worth noting that on more demanding text scenarios, higher transition rates were associated with better scores. This suggests that more frequent shifts in attention between these areas may have been the most effective strategy.

Overall, text display scenarios proved to be far more difficult for most participants as shown by performance data, reaction times and TLX workload ratings, especially at higher demand levels. Evidently, requiring participants to divide attention between the radar screen and the message windows to identify task demands had a detrimental effect on successful completion of the task. A depiction of the text display viewing strategy is provided in figure 7.

Modeling eye movements

In order to design better models of human performance and behavior, it is important to understand cognitive processes at the highest possible level of elaboration. There is no substitute for the vantage point into these processes that eye tracking can provide. As demonstrated in the current study, analysis of viewing percentages and transition rates can offer invaluable assistance in reconstructing the specific strategies used in a particular task, even as they change alongside task demands. If properly incorporated into the vision modules of the models examined in the AMBR project, this information could help to facilitate an even more robust level of accuracy. This research serves as a perfect example of the crucial role that eye tracking technology can play in piecing together cognition and supporting model development.

DISCUSSION II: WORKLOAD ANALYSIS

Interpreting Workload Relationships

ICA and the TLX

Analysis of the data reported here suggests that the relationship between ICA and the TLX may not be quite as simple as originally assumed. Ideally, there should be a positive relationship between subjective ratings and objective psycho-physiological measures of workload. They are both supposedly measuring the same characteristic of the task and both demonstrate a mean increase as level of demand increases, so it would seem fair to assume that these two measures should positively correlate with each other. For this reason, the observed negative association between ICA and TLX ratings seemed perplexing.

Further analysis of the *Taskload* variable shed some light onto the nature of this relationship. The *Taskload* provides the number of stimuli requiring response at any given point throughout the scenario. When averaged across the entire task, it can be conceptualized as an estimate of actual workload based on how task demands were managed. In other words, a high *Taskload* is indicative of an inability to execute necessary commands in a timely fashion. We tested the notion that the TLX may not have been measuring actual workload, but rather how well the cognitive effort put forth matched the actual workload required. The variable *ICA Load* was a ratio of *taskload* to ICA, created to address this possible relationship. By exploring this relationship between specific task demands and psycho-physiologically derived estimates of cognitive activity, we were able to derive an alternative explanation of how the TLX and ICA are related. Figure 8 below demonstrates the degree to which *ICA load* reflects TLX ratings at all three demand levels.

These data suggest that TLX ratings are a reflection of the degree to which cognitive activity keeps pace with task demands, especially at higher demand levels. The three correlations reflected in the figure are 0.36, 0.63, and 0.82 respectively. The latter two correlations are significantly different from zero at p<0.01 each.

Participants whose ICA remained relatively low while task demands were high reported these scenarios to be more difficult in their TLX ratings. Participants exhibiting high ICA while task demands were low reported less difficulty in their TLX ratings. To summarize, TLX ratings reflect not the level of workload for the task, but rather how adequate the cognitive effort expended was in complying with task demands. This trend is especially apparent as level of demand increases. Evidently, as the tasks become more difficult, the discrepancy between *Taskload* and ICA grows, and this ratio becomes even more highly associated with subjective workload ratings.

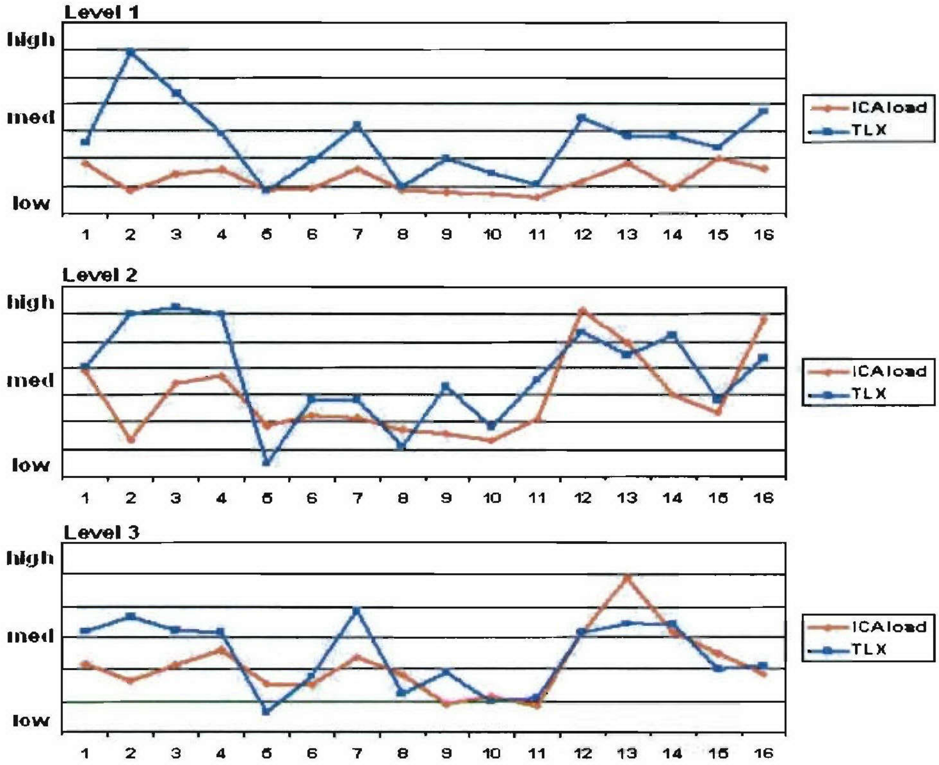


Figure 8. TLX and *ICA load* relationship for each subject at each demand level

A hierarchical multiple regression analysis was carried out in which TLX was predicted from taskload and the ICA. For the analysis, the dummy variables were entered first to account for between-subject variability. Next, *taskload* was entered separately, and finally *ICA* and *ICA_load* were entered together into the model. The three-stage model yielded the results shown in Table 6. Subject variability alone accounted for 35% of the variance. Adding *taskload* increased variance accounted for to 88%, and adding the ICA plus the interaction term of ICA_load increased variance accounted for to 92%. Both ICA and *ICA_load* had significant regression coefficients, indicating that both made unique and significant contributions to the model.

Table 6. Regression analysis predicting TLX from ICA and Taskload

Step	Variables in the Model	R^2	Significant R^2 Increase
1	all dummy variables	0.35	0.000
2	all dummy variables; taskload	0.88	0.000
3	all dummy variables; taskload; ICA and ICA_load	0.92	0.015

ACT-R workload and the TLX

A similar argument can be made about the relationship between ACT-R workload estimates and the TLX. The relationship between ACT-R predictions and TLX workload ratings corroborates the previously discussed evidence suggesting that the TLX may not be measuring workload per se. One would expect model workload predictions to match participant ratings of scenario difficulty. As with the ICA-TLX relationship, this was not the case. Higher model predictions of workload were not associated with higher TLX ratings. If the TLX is to be accepted as an accurate predictor of workload, it is imperative that it correlate significantly with measures that are free from subjective biases.

Following the same course of logic used to create the *ICA load* variable, we determined that it might be useful to analyze the degree to which ACT-R workload predictions kept pace with task demands. Since *Taskload* provides the best overall assessment of how effectively participants managed the task, it was again combined with the workload measure. *ACT-R load* was calculated by taking the ratio of *taskload* to ACT-R workload prediction for each participant on each scenario.

As in the case of the *ICA load*, *ACT-R load* correlates better with participant TLX ratings than did ACT-R estimates alone. Correlations were 0.41, 0.45, and 0.55 for levels 1, 2, and 3 respectively. At level 3, the correlation is statistically significant, p<.05. It seems that TLX ratings are at least partially dependent on the capability of the participant to complete tasks quickly and efficiently. Again, this suggests that subjective ratings are not measuring how hard you work, but rather how well your workload matches the amount of workload required to effectively manage the task. Evidence from both *ICA* and *ACT-R* lend credence to this contention.

A second regression analysis was carried out, substituting *ACT-R* and the *ACT-*R *load* variables for the *ICA* and *ICA load* variables. The analysis was again a three-stage model in which dummy variables were entered followed by *taskload* followed by the ACT-R variables. The results are given in Table 7. Unlike the ICA, the ACT-R additions did not make a significant contribution to the model.

Table 7. Regression analysis predicting TLX from ACT-R and Taskload

Step	Variables in the Model	R^2	Significant R^2 Increase
1	all dummy variables	0.35	0.000
2	all dummy variables; taskload	0.88	0.000
3	all dummy variables; taskload; ACT-R and ACT-R_load	0.89	0.755

ICA and ACT-R

The results presented here suggest that ACT-R and ICA both demonstrate capability to distinguish between demand levels. Overall estimates reveal the hypothesized trend that as demand level increases, both ICA and ACT-R predictions increase as well. However, the comparison of these two estimates throughout the task did not reveal a significant correlation. The two measures draw upon very different resources to assess the amount of workload in a given task. ACT-R calculates workload as a scaled ratio of time spent in critical tasks to time on the entire task. Critical tasks include any responses to onscreen stimuli as well as scanning the message windows and patrolling the radar screen in search of new action prompts that must be addressed. ICA, on the other hand, is based not on onscreen events or behaviors, but rather specific physiological reactions to cognitive activity sampled 250 times every second. For this reason, it is not entirely surprising that these two measures would be difficult to equate on a second-by-second basis.

The analysis of specific time intervals demonstrates the dissociation between ACT-R and ICA in that there does not seem to be a perceptible pattern in how these variables relate over the course of each scenario. Sometimes ICA is high when ACT-R predictions are high, yet other times the opposite is true. The degree to which these variables are correlated does not significantly impact participant performance, but it does seem to have an effect on subjective workload ratings. The more closely associated these two variables are across 30-second intervals, the easier participants rate the tasks to be. This suggests that, although these variables are not conducive to comparison over whole scenarios, their relationship to each other throughout the task may relate to subjective workload ratings.

The relationship between TLX ratings and the correlation between the ACT-R and ICA workload measures also hints at one of the difficulties in predicting behavior patterns at a fine grain of analysis. The easier participants rated the task, the more closely related the two workload measures were correlated. The ease of task in this case is reflective of scenarios where events that required a participant response occurred relatively less frequently than in other scenarios. In this case, as a participant was performing the task, an event, such as a plane requesting acceptance, would occur, and the participant would have time to respond to that event before another event requiring a response would occur. Under those circumstances, the behavior of participants is relatively constrained, i.e., respond to the lone event. However, in a scenario where there are multiple events that require a response, the behavior of the participants is relatively less constrained by the demands of the task. For instance, when there were multiple events requiring a response, the strategy participants used for prioritizing their responses could vary widely. They might choose to respond to events in the order they occurred, or they might choose to respond to events based on the number of penalty points associated with a delayed response. It is likely that the more the task constrains participant behavior, the more precisely their workload may be predicted by a computational model and inferred from psycho-physiological measures. Hence, the higher correlations between the ACT-R workload measure and the ICA measure found where participants rated the task as easier, is likely due in part to the fact that the participants' actions were more constrained at easier levels of the task. Further methodological problems in using this task to assess workload are discussed in the following section.

Overall Workload Assessment

Cognitive workload is a difficult construct to quantify and, perhaps even more so, to define. The short answer is that cognitive workload is the degree to which an operator's cognitive and perceptual capabilities are taxed during completion of a task. This definition is perfectly adequate, as long as there is agreement on how best to measure cognitive and perceptual capabilities. As demonstrated in the research presented here, this is not the case. The TLX, ICA, and ACT-R all provided estimates of cognitive workload using very different means which captured very different aspects of cognition and perception. Although all three measures distinguished between demand levels, ACT-R and ICA proved difficult to associate during each scenario, and the TLX demonstrated an unexpected negative relationship to both other measures. It is exceedingly optimistic to expect these three methods to be tapping the same cognitive reservoir. In fact, there was very little agreement among these measures in their individual quantifications of workload. This is not to say that one method is correct and the other two are faulty. The most likely explanation is that these three tools are measuring different facets of the broad concept of cognitive workload.

The NASA Taskload Index (TLX)

The strengths of the TLX include its ease of administration and scoring, as well as its overall correlation with task performance. For research requiring a practical method of obtaining workload estimates that is not concerned with problems of subjectivity, the TLX is an adequate tool. In the AMBR context, the TLX was successful in identifying differences between the three demand levels. However, our analysis suggests that TLX may not have been measuring workload in itself. The extremely high correlation between TLX ratings and *ICA load* and, to a lesser extent, *ACT-R load* suggests that when people rate the level of workload in a given scenario, they are not rating the overall cognitive effort, but rather how well the amount of effort put forth met the demands of the task.

It may seem like an insignificant distinction, but, when considered in terms of overall task analysis, it is important. A participant may put forth very little cognitive effort on an easy task and rate the task as high in workload because the minimal effort put forth did not keep pace with task demands. On the other hand, a participant may expend a great deal of cognitive effort on a difficult scenario, such that the demands of the task are managed very effectively. In this situation, this research suggests that the TLX rating would be low. Clearly, in both of these cases the TLX characterization of workload is misleading. It is not the amount of cognitive effort that is being measured, but how adequately that effort allows the participant to successfully navigate the task environment. When using the TLX to rate the level of cognitive workload, it is useful to keep this distinction in mind.

ACT-R workload predictions

The ACT-R method of workload assessment avoids the pitfalls of subjectivity demonstrated by the TLX. Nonetheless, as a predictive measure, it suffers from an opposite limitation. By virtue of being a predictive model, ACT-R estimates workload without the benefit of any information from the subjects whose workload it attempts to predict. Beyond the behavioral data regarding participant responses, the model is left to predict cognitive workload based on assumptions regarding where the participant should be looking, how many items they should be storing in memory and, in a broader sense, how cognitively demanding the overall task is.

Not surprisingly, the ACT-R estimates were correlated with the *taskload* variable at 10 second intervals. Correlations between these two variables for individual subjects reached as high as .50 and averaged a significant .26. *Taskload* was calculated each second by summing the number of critical tasks to be completed. In many ways this correlation between ACT-R estimates and *taskload* at discrete intervals may be regarded as a success in that ACT-R was able to predict workload associated with on screen events and task requirements throughout scenarios. Based on this, we contend that ACT-R predictions accurately portrayed workload as it relates to task management.

However, as in the case of the TLX, the argument can be made that this measurement, useful as it may be in performance analysis, does not truly predict cognitive workload in itself. ACT-R does an excellent job of plotting operator activity and assigning a level of workload to each of those activities, but the cognitive weight given to each activity is an assumption, as is the place in time that cognitive activity takes place. This is not a shortcoming of the model, but rather an unavoidable consequence of working with a complex task such as air traffic control.

There are two practical drawbacks to predicting workload on this task. First of all, there are extended segments of time during each scenario in which there are no stimuli requiring attention. During these periods, the model predicts workload at a predetermined baseline. In many cases participants use this time to search the screen for the coming events, evaluate what has already been done or think about other things that may not be related to the task at all. To characterize these very different cognitive activities with the same workload is misleading. Secondly, in periods of high onscreen activity, assumptions about the level of workload may be misguided. There is no way of knowing when multitasking occurs and when individual tasks or parts of tasks are handled piecemeal. Some actions may be completed without any appreciable increase in workload, while others require a substantial increase. The current computation from which ACT-R derives workload may be adequate for simpler tasks in which actions are more closely tied to cognitions, but to suggest that workload can be predicted in these scenarios by simply tracing periods of time in which critical tasks are undertaken is like suggesting that cognition is the same as behavior. It is important to make this distinction when interpreting the ACT-R workload estimates; they are extremely accurate in describing task management, but they inevitably underestimate the complexities of human cognition on a task such as air traffic control.

The Index of Cognitive Activity (ICA)

The picture of cognitive activity portrayed by the ICA is quite different from the other two measures. ICA presents an overall level of cognitive activity that does not differ as dramatically between levels. ICA is low at all three levels, indicating that the specific tasks are not cognitively demanding once learned. Despite this fact, ICA is capable of distinguishing between demand levels and providing information in fine detail regarding when and where cognitive activity occurs. Participants exhibiting a higher ICA spent more time viewing the message windows and less on the radar screen. This suggests that more carefully monitoring the aircraft movements as opposed to the text was associated with limiting the amount of cognitive effort exerted. More specifically, participants with a high ICA experienced a greater portion of cognitive workload when viewing the outgoing messages and the region between the message windows and radar screen. This suggests that participants experiencing higher workload were more cognitively active while searching the messages for information on outgoing aircraft and when transitioning between the message windows and the radar screen.

It is also interesting to note that at the three demand levels, the distribution of cognitive activity in most regions was nearly identical. The exceptions were the amount of cognitive activity in the *inner square*, which decreased as level increased, and the amount of cognitive activity in the *east exterior*, which increased with increasing demand level. ICA in all other regions remained remarkably constant. This demonstrates the capability of the ICA to detect not only changes in cognitive activity, but the specific areas in which these changes occur.

Another finding from the ICA was that score was not significantly associated with high cognitive workload. This makes sense in light of the limitations of this task. There are two reasons that a participant may perform poorly on the ATC task: either they are cognitively overloaded, or they are not sufficiently cognitively engaged. Based on this, a high level of cognitive workload may just as easily be associated with good performance as bad performance. TLX and ACT-R estimates do not reflect this duality. This may be part of the reason that ICA does not correlate well with the other workload measures.

ICA also revealed a training effect, such that participants who learned color scenarios first exhibited significantly higher ICA on all scenarios. This distinction was apparent in the decreased performance of participants who learned in this order. Evidently, the manner in which participants are trained on a task such as this has an effect on the level of cognitive workload experienced during future completion of the task. Only ICA provides the raw cognitive data capable of assessing these aspects of the AMBR task.

Post-experiment analysis of the task revealed that a participant's responses to each stimulus do not actually require substantial cognitive effort. Once those responses have been learned, reacting to each stimulus becomes more or less an automated process. When an AC requests acceptance, for example, there is not a choice that needs to be made. It is as simple as identifying the stimulus and clicking buttons in a learned sequence. In the higher demand level scenarios, there is not an increase in the cognitive requirements. In fact, the type and number of actions to be completed are exactly the same. The added complication is time pressure, which is not the same as adding cognitive workload. Transferring an AC at demand level 1 and level 3 require the same cognitive effort. The only difference is the increased pressure to complete the action quickly and move on to the next one.

This is not to say that cognitive activity can not be associated with particular aspects of the task. As demonstrated in figure 9, ICA can be used along with eye movement information to ascertain where and when cognitive activity occurs. As displayed in this picture, specific events can often be associated with these cognitive bursts.

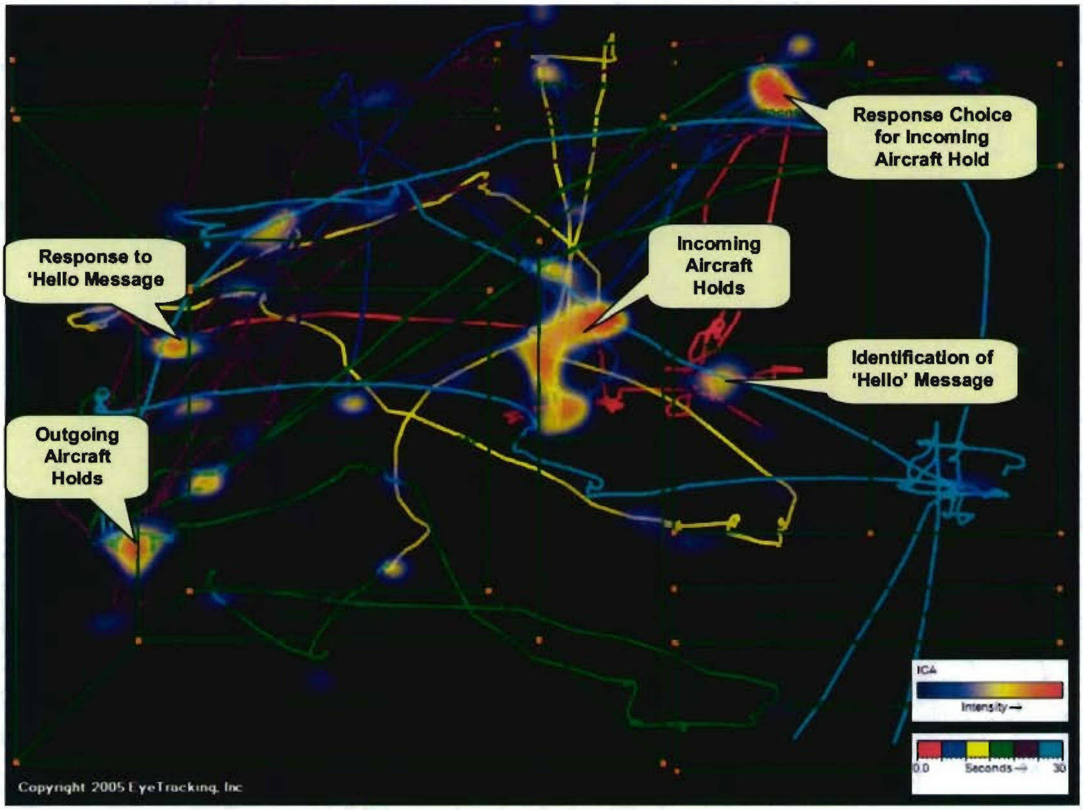


Figure 9. Plot of participant 13 gaze and ICA for 30-seconds of scenario B6 Text.

CONCLUSIONS

This research provided the opportunity to both analyze eye movements on the AMBR task and compare psychophysiological estimates of cognitive workload with computational model predictions. Eye tracking revealed specific gaze and eye movement patterns that were associated with different display types and demand levels. This information could be utilized to improve the accuracy with which models simulate human performance.

The analyses from this study elucidate important distinctions between the three workload measures. The NASA TLX seems to be measuring how well the amount of cognitive effort expended by a participant met the demands of the task. The ACT-R workload predictions accurately match workload to on-screen task management and behavior. Finally, the ICA provides an estimate of the raw cognitive activity expended while completing a task. All of these yield important information for assessing the AMBR task and entail their own unique strengths and weaknesses. The degree to which any of these measures independently examine the broad construct of cognitive workload depends on how one defines the term.

REFERENCES

Deutsch, S., & Cramer, J. (1998). Omar Human Performance Modeling in a Decision Support Experiment. In *proceedings of 42nd Annual Meeting of the Human Factors and Ergonomics Society*, Chicago, IL.

Gray, W. (2000). *Summary of the AMBR Expert Panel Report.* http://www.mesa.afmc.af.mil/AMBR/AMBR1_Gray.ppt.

Hart, S. & Staveland, L. (1988). Development of NASA-TLX (*Taskload* Index): Results of empirical and theoretical research. In P.A. Hancock & N. Meshkati (Eds.) *Human Mental Workload.* Elsevier.

Hornof, A., & Halverson, T. (2003). Cognitive strategies and eye movements for searching hierarchical computer displays. In *proceedings of the Conference on Human Factors in Computing Systems, ACM*, 249–256.

Lebiere, C. (2001). A theory-based model of cognitive workload and its applications. In *Proceedings of the 2001 Interservice/Industry Training, Simulation and Education Conference* (I/ITSEC). Arlington, VA: NDIA.

Lebiere, C, Anderson, J.R., & Bothell, D. (2001). Multi-tasking and cognitive workload in an ACT-R model of a simplified air traffic control task. In Proceedings of the Tenth Conference on Computer Generated Forces and Behavior Representation. Norfolk, VA.

MacMillan, J., Deutsch, S., & Young, M. (1997). A Comparison of Alternatives for Automated Decision Support in a Multi-task Environment. *In Proceedings of the 41st Annual Meeting of the Human Factors and Ergonomics Society,* 190- 194.

Marshall, S. (2005) Assessing Cognitive Engagement and Cognitive State from Eye Metrics. *HCII2005 Conference Proceedings*. Las Vegas, NV:CD-ROM.

Marshall, S. (2000). *U.S. Patent No. 6,090,051.* Washington, DC. Patent & Trademark Office.

Rehling, J., Lovett, M., Lebiere, C., Reder, L. M., & Demiral, B. (2004) Modeling complex tasks: An individual difference approach. In proceedings of the 26th Annual Conference of the Cognitive Science Society (pp. 1137-1142) . August 4-7, Chicago, USA

Salvucci, D. (2005). A Multitasking General Executive for Compound Continuous Tasks. *Cognitive Science, 29,* 457-492.

Schvaneveldt, R., Reid, G. B., Gomez, R. L., & Rice, S. (1998). Modeling Mental Workload. *Cognitive Technology*, 3, 19-31.

Son, I., Guhe, M., Gray, W., Yazici, B., & Schoelles, M. (2005) Human Performance assessment using fNIR. In: *Proceedings of SPIE 5797: Biomonitoring for Physiological and Cognitive Performance during Military Operations*, 158–169. 31 March–1 April 2005, Orlando, FL.

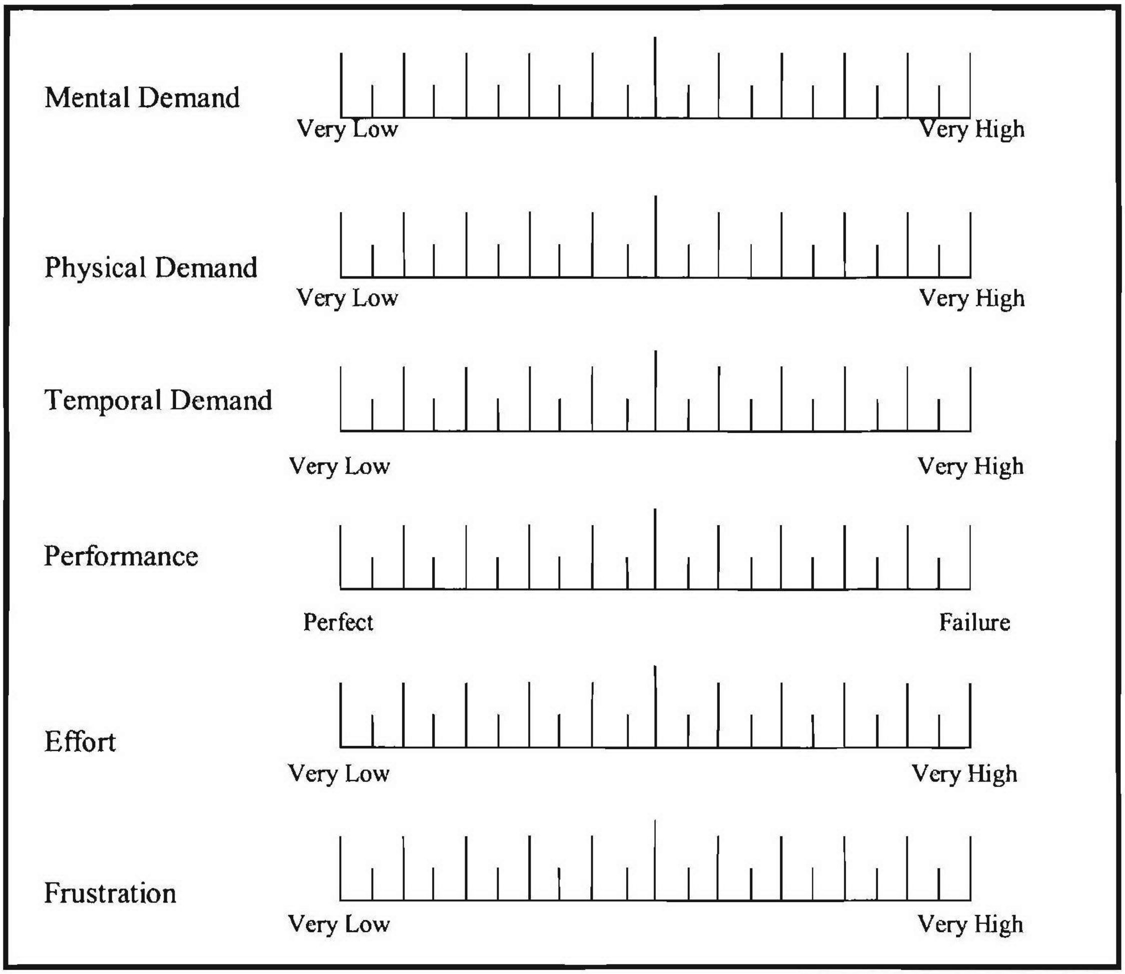
Tenney, Y., & Spector, S. (2001). Comparisons of HBR Models with Human-in-the-loop Performance in a Simplified Air Traffic Control Simulation with and without HLA Protocols: Task Simulation, Human Data and Results. In *Proceedings of the 10th Conference on Computer-Generated Forces and Behavior Representation*, IEEE/ITCMS, Piscataway, NJ, 15-26.

FOOTNOTES

[1] The authors acknowledge the contribution of Christian Lebiere and Michael Fleetwood of Micro Analysis and Design, whose assistance in describing the ACT-R model, providing workload predictions and assisting in interpreting data was essential to the completion of this project.

APPENDIX A: The NASA TLX

Subject No.__________________
Condition Code_______________



Title	Endpoints	Description
Mental Demand	Very Low/Very High	How mentally demanding was the task?
Physical Demand	Very Low/Very High	How physically demanding was the task?
Temporal Demand	Very Low/Very High	How hurried or rushed was the pace of the task?
Performance	Perfect/Failure	How successful were you in accomplishing what you were supposed to do?
Effort	Very Low/Very High	How hard did you have to work to reach your level of Performance?
Frustration	Very Low/Very High	How insecure, discouraged, irritated or annoyed were you?

APPENDIX B: Modifications for the Current ACT-R Model

The model's declarative memory is based on four chunk types, which were defined using the **chunk-type** command. They consist of the name of the chunk type and the associated slots. In the original model there were two chunk types associated with top-level goals, however, because we were focusing solely on the text condition, only one of those chunk types was used in the current effort, **text-goal**. The other three chunk types relating to goal structure were left unaltered, **scan-text**, **scan-screen**, and **process**. The chunks related to these three chunk types and defined by the add-dm command are simply symbols used in other chunks (which the system would define by default) and the initial goal text condition. They and their associated procedural knowledge will be described in detail in the rest of this section. The productions that apply to each goal type will be listed in a table using an informal English description that is meant to capture their function without obscuring syntactic details. Production names are in bold while words in italics correspond to production variables and words in bold within the production text correspond to specific chunks (constants).

In the original model, the **text-goal** unit task directed attention to specific areas of the screen in order to scan for events that required an action. The current model follows that same basic pattern, however, the method for determining which screen area to examine was altered. In the original model, attention was directed to the different screen areas in a sequential fashion. In the current model, attention is directed to the different screen areas randomly, until the model approaches the time when the next participant action (gleaned from the participant's experiment log) was initiated. When the model time is within 0.3 seconds of the next participant-initiated event, the model directs its attention to the portion of the window where the event corresponding to the participant's action is located. For instance, if the next event in the participant log is to "Welcome" an airplane and it occurred at 53.6 seconds, then the first time after 53.3 seconds of model simulation time that the model directs its attention to a portion of the screen to scan for an event, its attention will be directed at the portion of the screen where "Welcome" messages are located. (The parameter of 0.3 seconds was chosen because several additional productions must fire before the model actually initiates the user action—generally clicking a button—and initiating the processing of the text window 0.3 seconds in advance allowed the model to click on the button at very nearly the same time as the participant.) This sequence is accomplished through three additional productions, which are presented in Table B.1.

Table B.2 presents the productions for the unit task **scan-text** responsible for scanning a text window. In the original model, **scan-text** goals started scanning at the bottom of the screen. The production **find-flush-message** scanned upward from the current position (initially bottom) to find the next message that is flush against the left side of the window, indicating a message from an aircraft or another controller requesting action. In the current model,

find-flush-message does not scan from the bottom, but simply finds the message corresponding to the next user-initiated action. (It is assumed that the user found such a message in order to initiate his/her action.) If the ACT-R simulated time is not within 0.3 seconds of the next action, the production **no-flush-message** pops the goal, which returns control to the **text-goal** unit-task. In the original model, if a message was found requesting action, the model then tried to determine whether that action had already been completed. The production **memory-for-message** searched declarative memory for a chunk recording the completion of a process goal for the task and aircraft indicated by the message. Similarly, if the memory retrieval failed the production **message-reply** scanned down the text window from the current message for an indented message containing the acknowledgment message that would have resulted from taking that action. If either a memory or a message indicating completion of the action was found, the goal was popped. However, in the current system, in order for the model to trace the actions of the participant, these two productions were eliminated. Occasionally, participants made the mistake of responding to events multiple times, and in these cases, the model was constrained to perform the same actions as a participant, regardless of the need for the action. After the message corresponding to the next user action was found, the production **subgoal-message-task** pushed a sub-goal to perform that action and cleared the goal to allow further scanning to take place when that unit task was completed.

Window-to-window-time-near-text

IF the *goal* is of type text-goal and the next participant event is within 0.3 seconds and the next event is a message event

THEN push a subgoal to scan the text area in the next event window

Window-to-window-time-near-screen

IF the *goal* is of type text-goal and the next participant event is within 0.3 seconds and the next event is a screen event

THEN push a subgoal to scan the screen area in the next event window

Window-to-window-time-too-far

IF the *goal* is of type text-goal and the next participant event is NOT within 0.3 seconds

THEN push a subgoal to scan a random text area

Table B.1: Productions applicable to the unit task **text-goal**

Two additional productions that were removed from the model for reasons related to model tracing were the productions **detect-onset-text** and **focus-onset-text**, which provided the capacity to detecting the onset of a new message in other text windows (not currently attended) and record in the current goal to focus attention to that window as soon as the current message had been processed. Again, the model tracing system was developed to constrain the model to follow a participant's actions, regardless of the context of the task. Hence, even if multiple messages appeared that required an action on the part of the participant, if the participant made no response, then the model did not either.

Find-flush-message

IF the *goal* is of type scan-text of area *window* and no aircraft is currently selected and *message* is the message corresponding to the next user action and the next participant event is within 0.3 seconds

THEN note the *task*, *aircraft* and *controller* in *message*

No-flush-message

IF the *goal* is of type scan-text and no aircraft is currently selected and the next participant event is NOT within 0.3 seconds

THEN pop the current goal

Subgoal-message-task

IF the *goal* is of type scan-text with task *task*, aircraft *aircraft* and controller *controller*

THEN clear *goal* and push *subgoal* to process task *task* on aircraft *aircraft* with controller *controller*

Table B.2: Productions applicable to the unit task **scan-text**

Table B.3 presents the productions for the unit task **scan-screen** responsible for scanning the radar screen, more specifically the area between the green and yellow lines in which exiting aircraft that need to be transferred can be detected. Because of the similarity between the two unit tasks, both of which consists in scanning a screen area to detect events that require actions, the set of productions for the unit task scan-screen is quite similar to those for the unit task **scan-text**. **Scan-for-transfer** no longer scans the radar area for exiting aircraft, as it did in the original model, but rather locates the aircraft corresponding to the next action by the participant. **Memory-for-transfer** and **trace-of-transfer** were removed from the model since the action must be completed regardless of whether it had been done before. **Subgoal-transfer** pushes a sub-goal to transfer the aircraft. If no more exiting aircraft can be detected, **scan-done** pops the goal. The message onset detection productions **detect-onset-screen** and **focus-onset-screen**, were removed similar to their counterparts in unit task **scan-text**.

One additional detection production was removed in the current model. **Detect-red** detected a red aircraft indicating a holding violation. However, the model would only respond to that aircraft if it was the next aircraft acted upon by a participant, regardless of its color.

Scan-for-transfer

IF the *goal* is of type scan-screen and no aircraft is currently selected and *aircraft* is the aircraft corresponding to the next user action

THEN note *aircraft* with its *position* and associated *controller*

Scan-done

IF the *goal* is of type scan-screen and no aircraft is currently selected

THEN pop *goal*

Subgoal-transfer

IF the *goal* is of type scan-screen with current *aircraft* in *position* with *controller*

THEN clear *goal* and push *subgoal* to process transfer on *aircraft* in *position* with *controller*

Table B.3: Productions applicable to the unit task **scan-screen**

Table B.4 presents the productions for the unit task process responsible for actually processing an action request through a sequence of button clicks and mouse selections. The first action to perform is to click the button on the right side of the screen corresponding to the requested action. The production **answer-speed-request** fires if the next participant action was to respond to a speed request. If so, the production pushes the button corresponding to the participant's response to the speed request (accept or reject). The production **answer-other-requests** pushes the corresponding button for all other actions. The next action is to select the aircraft. However, in some conditions (e.g. responding to a text message) the location of the aircraft is not yet known and the aircraft will have to be located first. Three productions that could have fired at this point in the original model, **memory-for-position**, **find-position-inner**, **find-position-between** and **find-position-outer** were removed, as the current model assumed that the participant was able to successfully locate the aircraft's position. Next, the target was selecting by the production **click-target**. The production **click-controller** then selected the external controller associated to the aircraft, unless preempted by productions **skip-speed-change-controller** and **skip-welcome-controller** that explicitly skip that step for the speed change and welcome actions respectively. The **click-send** production was responsible for clicking the send button and popping the goal. However, click-send could not fire until the ACT-R model time had reached the time at which the participant clicked the send button for the action. If that time had not yet arrived, then the production **wait-to-click-send** would fire repeatedly until it did. This ensured that the total time for the sequence of actions, from selecting the button corresponding to the action to clicking the send button was very nearly reproduced by the model (to within 0.3

seconds). Two other relevant productions were also added to the current model, **click-cancel** and **skip-send-cancel**, which fired in lieu of **click-send** if the participant clicked the cancel button instead of send, or, if the participant did not click either button before moving onto the next action, the **skip-send-cancel** production would fire.

Table B.4: Productions applicable to the unit task **process**

Answer-speed-request

IF the *goal* is of type process with action speed-change for *aircraft* in *position* and step select

THEN determine participant's action for request and push button corresponding to accept-reject decision and note that the step is now target

Answer-other-requests

IF the *goal* is of type process with *action* and step select

THEN push button corresponding to *action* and note that the step is now target

Click-target

IF the *goal* is of type process with *aircraft* in *position* and step target

THEN select *aircraft* in *position* and update step to controller

Skip-speed-change-controller

IF the *goal* is of type process with action speed-change and step controller

THEN update step to send

Skip-welcome-controller

IF the *goal* is of type process with action welcome and step controller

THEN update step to send

Click-controller

IF the *goal* is of type process with *aircraft* step controller

THEN select *controller* associated with *aircraft* and update step to send

Click-send

IF the *goal* is of type process with step send and the participant clicked the send button and the current time is after when the participant clicked the button

THEN push button send and pop *goal*

Click-Cancel

IF the *goal* is of type process with step send and the participant clicked the cancel button and the current time is after when the participant clicked the button

THEN push button Cancel and pop *goal*

Wait-to-click-send

IF the *goal* is of type process with step send and the participant clicked the send or cancel button and the current time is NOT after when the participant clicked the button

THEN do not change the goal

Skip-send-cancel

IF the *goal* is of type process with step send and the participant did not click the send or cancel button

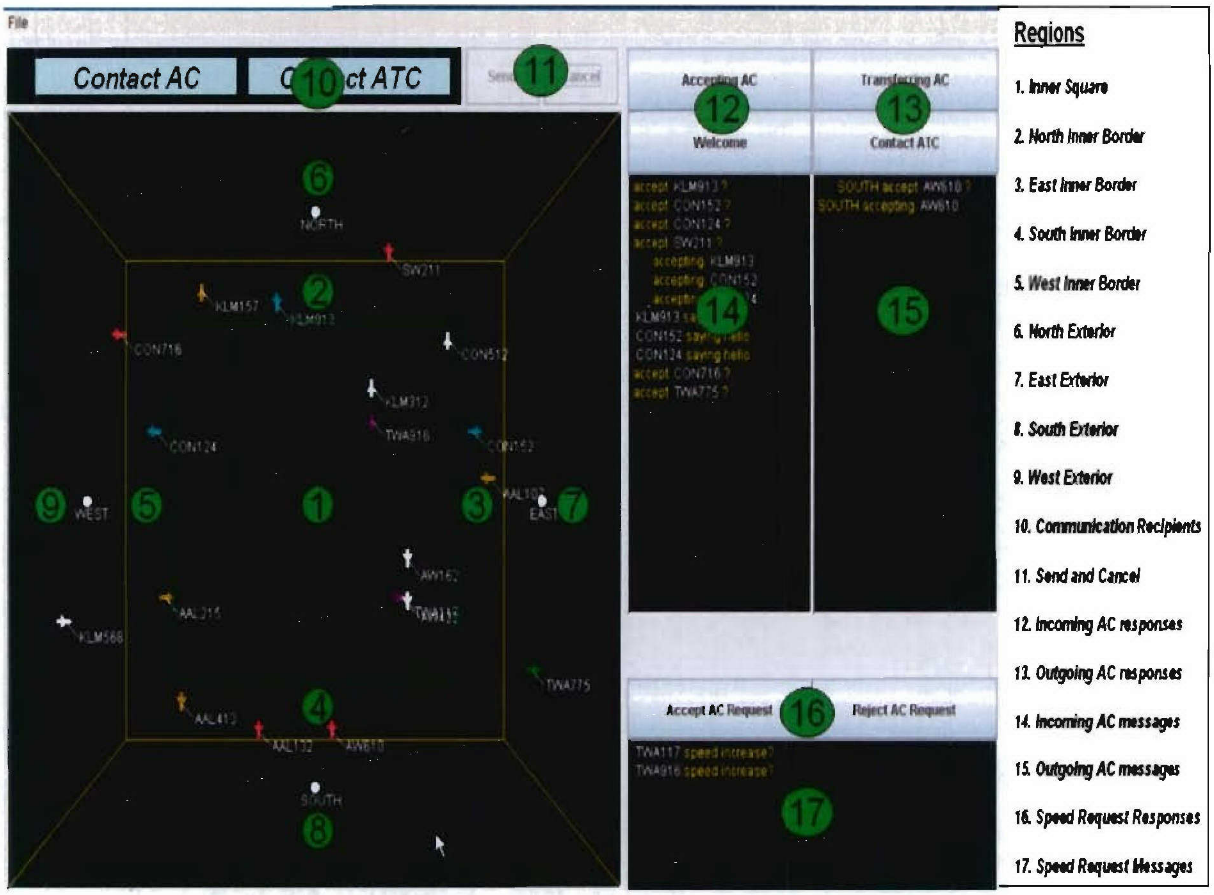
THEN pop the *goal*

Table B.4: Productions applicable to the unit task process (continued)

The final part of the model concerns the code at the top of the model that is used to compute the workload estimates. While ACT-R has traditionally shied away from such meta-awareness measures and concentrated on matching directly measurable data such as external actions, response times and eye movements, it is by no means incapable of doing so. For the purpose of this model, we proposed a measure of cognitive workload in ACT-R grounded in the central concept of unit task. Workload is defined as the ratio of time spent in two unit tasks to the total time spent on task. The two different unit tasks, a scanning task and critical tasks, were weighted differently in the workload calculation. Critical unit tasks were defined as tasks that involve actions, such as the **process** goal that involves handling an event with 3 or 4 mouse clicks and the processing that was required to accomplish such tasks. These critical tasks were given a weighting of 1.0 in the workload calculation. Hence, if the model were involved in a critical task for an entire relevant time frame (e.g. five seconds), then the workload calculation would be 1.0. The scanning task was defined as when the model was scanning the screen for text messages but was not processing or responding to an event. It was given a weighting of 0.34 in the workload calculation. This number, 0.34, was derived from the pupil data that the workload measure was compared to. It was the average ICA for the timeframe from 10.0 seconds to 20.0 seconds of all the participants in the level 3 demand condition (longest time frame) of the text condition. This timeframe was used as it was a period when no events occurred that required a response from a participant; hence it was a time frame when they were all scanning the screen in search of the next relevant event to respond to.

APPENDIX C: Task Regions

Air traffic controller display with defined regions on a color scenario



accept KLM913 ?
accept CON152 ?
accept CON124 ?
accept SW211 ?
accepting KLM913
accepting CON152
accepting CON124
KLM913 saying hello
CON152 saying hello
CON124 saying hello
accept CON716 ?
accept TWA775 ?

SOUTH accept AW610 ?
SOUTH accepting AW610

TWA117 speed increase?
TWA916 speed increase?

Regions

1. Inner Square
2. North Inner Border
3. East Inner Border
4. South Inner Border
5. West Inner Border
6. North Exterior
7. East Exterior
8. South Exterior
9. West Exterior
10. Communication Recipients
11. Send and Cancel
12. Incoming AC responses
13. Outgoing AC responses
14. Incoming AC messages
15. Outgoing AC messages
16. Speed Request Responses
17. Speed Request Messages